Social Functions of Machine Emotional Expressions

Celso de Melo, Jonathan Gratch, Stacy Marsella, Catherine Pelachaud

▶ **To cite this version:**

Celso de Melo, Jonathan Gratch, Stacy Marsella, Catherine Pelachaud. Social Functions of Machine Emotional Expressions. Proceedings of the IEEE, 2023, 111 (10), pp.1382-1397. 10.1109/JPROC.2023.3261137 . hal-04293243

HAL Id: hal-04293243
https://hal.science/hal-04293243

Submitted on 21 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Social Functions of Machine Emotional Expressions

Celso M. de Melo, Jonathan Gratch, Stacy Marsella, Catherine Pelachaud

***Abstract*—Virtual humans and social robots frequently generate behaviors that human observers naturally see as expressing emotion. In this review article, we highlight that these expressions can have important benefits for human-machine interaction. We first summarize the psychological findings on how emotional expressions achieve important social functions in human relationships and highlight that artificial emotional expressions can serve analogous functions in human-machine interaction. We then review computational methods for determining what expressions make sense to generate within the context of an interaction and how to realize those expressions across multiple modalities such as facial expressions, voice, language and touch. The use of synthetic expressions raises a number of ethical concerns and we conclude with a discussion of principles to achieve the benefits of machine emotion in ethical ways.**

***Index Terms*—Emotional Expression, Social Robots, Virtual Humans, Social Functions, Affective Computing.**

I. Why Should AI Express Emotion?

THE argument for expressing emotion in machines is simple: there is increasing evidence that emotional expressions serve essential social functions among humans [1], [2], including the facilitation of efficient and non-threatening communication [3], [4], telegraphing intentions and mental state [5], building trust [6], promoting fairness and cooperation [7], [8], shaping everyday decision making [9] and helping others to regulate their own emotions [10]. Therefore, to the extent that it is possible to realize these functions in human-machine interaction [11]–[15], machines have the potential to be more successful in teaming, building trust, and promoting cooperation with humans by appropriately expressing emotion (Figure 1). Here we present a critical review of the theoretical foundations and empirical evidence supporting this argument, emphasizing computational approaches, practical applications, and opportunities and challenges for the future.

A. Social Function, Not Internal Emotion

In reviewing the role of emotional expressions, we emphasize that expressions of emotion are not the same as internal emotional mechanisms or feelings. Some in affective computing have argued that any sufficiently sophisticated machine must implement internal mechanisms analogous to human emotion [21], and thus expressions might directly convey the "true feelings" of the machine. While this can serve as the basis for generating coherent expressions in machines (see Section III), this is not necessarily how emotional expressions "work" in human social interactions. As reviewed in Section II, expressions may reveal underlying emotional state, but people routinely mask or regulate their expressions for various social purposes (e.g., masking anger with a smile out of politeness) or deliberately produce expressions with the intent to influence their partners.

These differences are crucial for those interested in how people generate expressions, but they may be less important for understanding the impact of these expressions on observers. If someone expresses sadness, whether the individual is truly experiencing sadness or whether the individual is seeking consolation from others, the social impact of the expression may be the same. This is especially relevant for the case of emotional machines, as the expression can be trivially dissociated from the machine's internal state. On the one hand, this raises important ethical issues about the appropriateness of emotional expression in machines (see Section V); on the other hand, it defines a clear purpose for the expression of emotion in machines: to achieve the desired social effect on human users.

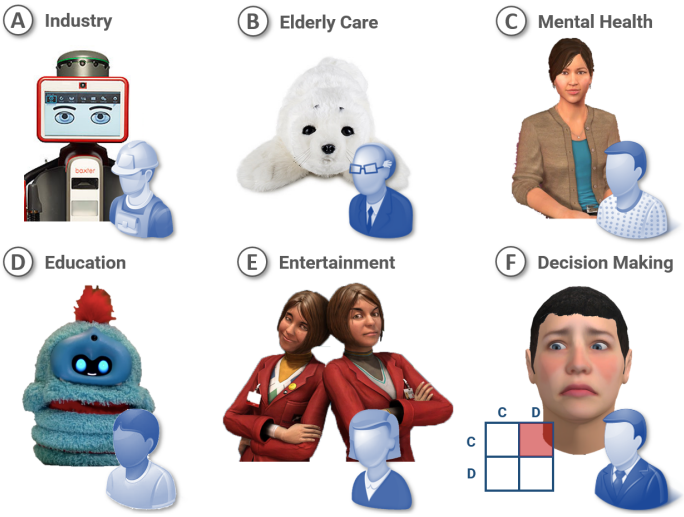

Fig. 1. Expression of emotion in machines is central in many social applications: A, Baxter supports social interaction with co-workers in industry settings [16]; B, Engaging with Paro has led to positive effects on elderly care [17]; C, The Simsensei kiosk demonstrates the potential of using virtual humans to assess sensitive mental health issues [18]; D, Multimodal expression in Tega contributed to learning success with children [19]; E, Expressive machines can also be used for entertainment applications, such as museum guides [20]; F, Expression of emotion can shape cooperation when engaging in social decision making with humans [5].

Manuscript received X X, 2022; revised X X, 2022.

This work was supported in part by ... *(Corresponding author: Celso M. de Melo.)*

Celso M. de Melo is with the DEVCOM US Army Research Laboratory (ARL), 2800 Powder Mill Rd, Adelphi, MD 20783 (email: celso.miguel.de.melo@gmail.com).

Jonathan Gratch is with the USC Institute for Creative Technologies, 12015 E Waterfront Dr, Los Angeles, CA 90094 (email: gratch@ict.usc.edu).

Stacy Marsella is with the Northeastern University, 360 Huntington Ave, Boston, MA 02115 (email: marsella@ccs.neu.edu).

Catherine Pelachaud is with the CNRS-ISIR, Sorbonne University, 4 place Jussieu, 75005 Paris, France (email: catherine.pelachaud@sorbonne-universite.fr).

B. Why Emotional Expressions?

While emotional expressions serve crucial social functions in human relationships, couldn't machines achieve these in other ways? For example, anger and guilt communicate and reinforce social norms [5], [8], but reputation mechanisms accomplish the same result without appealing to emotion [22]. Indeed, as noted by Herbert Simon in the early days of AI, machines need not replicate the exact workings of human cognition to realize its abstract function [23] and, thus, an analysis of the social functions of emotion can be fertile ground for hypothesizing novel abstract mechanisms. Nonetheless, there can be clear benefits to machines that generate behavior that observers would be comfortable labeling as *emotional*, and we focus on such approaches in our review.

In natural communication [24], [25], people efficiently use not only speech but gestures, expressions, and intonation to communicate propositional and non-propositional information efficiently and in a manner that is natural and familiar to humans and, ultimately, contributes to increased rapport, trust, and cooperation [5], [26], [27]. This has motivated computer scientists to consider the role of nonverbal behavior for the design of artificially intelligent machines. Justine Cassell, one of the early pioneers, highlighted that designers have many choices for a machine that is represented through its user interface, and argued cogently that human-like interfaces could be particularly effective in those cases where social collaborative behavior is key [28].

Emotional expressions merit special attention due to their established role in shaping interactions between people, thereby suggesting its utility for human-machine interaction. Empirical research reinforces the potential benefits of this design choice [29]. Emotional expressions lead users to perceive virtual and robotic machines as social actors [28], [30], [31], which can increase rapport and trust [26]. Emotional expressions have shown benefit across industrial collaboration [16], elderly companionship [32], education [19], [33], [34], and therapy [18]. Furthermore, human-machine studies have replicated behavioral sciences findings where expressions in machines communicated mental states to others and shaped decision making in social dilemmas [5], [35], [36] and negotiation [37]. In commercial settings, there is also interest in endowing voice assistants, such as Alexa and Google Assistant, with the ability to convey emotion [38]. However, a clear understanding of the functions of such expressions will be essential to ensure the success of all these applications.

Finally, in arguing for the importance of emotional expressions, we are not claiming that machines must exactly replicate how people express emotion in natural conversations for expressions to achieve their social function. For example, actors express emotion quite differently but, through exaggeration and avoiding emotion regulation, they enhance the communicative experience [39]. Similarly, within our scope, we allow machines that differ considerably from human form but focus on techniques that produce behaviors that people would consistently label as expressing an emotion such as Star Wars' R2-D2 expressions [40]. Accordingly, experimental studies suggest that even simple emotion rules [5] and non-anthropomorphic forms [7], [41], [42] of expression can successfully shape user experience.

C. Critical Review Scope

Although emotion plays many potential roles in human-machine interaction, our review focuses on emotional expressions generated by machines and the impact of these expressions on the people that interact with these machines. Thus, while machines may benefit from recognizing and interpreting expressions of people [43], this is outside our scope. Similarly, though machines might benefit from something like an internal model of emotion to inform their own decision-making and behavior preparation [15], [21], this is again outside our scope except to the extent that such models influence how observers interpret and respond to the machine's expressions. Finally, reflecting the current state of the field, we focus the review on individual and dyadic interaction aspects of the social function of emotion in machines; but see Section VI for a discussion of analysis at the group and cultural levels.

II. Theoretical Foundations

A. Theories of Emotion & Social Functions of Emotion

What is emotion? Before reviewing *emotional* expression in machines, we must clarify the term *emotion*. Following Klaus Scherer, we define emotion, in humans, as an episode of synchronized changes in the body and mind of a person in response to the evaluation of an event as being relevant to a significant goal or concern (see [44], p. 697). These include changes in most or all of a set of components including cognitive (appraisal of the event), bodily (changes to the core and peripheral nervous system and endocrine system), motivational (a resulting coping or action tendency), motor (including vocal and facial expressions) and subjective components (self-reported emotional feelings). As such, emotions are event-focused (as they arise in response to an appraisal of a specific event), and are relatively intense and short in duration compared with other phenomena studied in the affective computing community, and tend to signal immediate changes in behavior relevant to addressing the eliciting event. In contrast, moods or chronic states like depression tend to have long duration and experiencers of the mood may not be able to report a specific event that triggered the feeling.

As emotion tends to coordinate activity across components, assigning a label of anger to an individual is short-hand for a number of interrelated inferences: they are appraising an event as negative but controllable, their metabolism is increasing [45], they are likely to lash out [46], they will probably express anger in face or voice, and they will self-report they are angry. In this sense, labeling an expression as emotion can be seen as applying a schema or heuristic to help interpret latent states in the individual showing this expression, and serve as a guide for how to respond. Thus, in the context of human-machine interaction, we argue, if a person assigns a label of anger to a machine, they are essentially forming expectations about the machine's goals, subsequent behavior, and how it will answer if asked how it feels.

Recent debate in emotion psychology has centered around the right ontology for labeling emotional expressions and if expressions, thus labeled, truly predict appraisals, bodily changes, action tendencies and feelings. Affective computing has been heavily influenced by Paul Ekman's basic emotion theory which argues that the components of emotion are organized by six or seven "basic" emotion circuits [47]. In the strong version of this theory, components are tightly linked: if an individual shows anger, it can be inferred, even without regard for context, that the individual is experiencing anger, which in turn, suggests how they are likely to next respond. Indeed, a recent survey by the Association for the Advancement of Affective Computing indicated almost half of commercial products identified realize this strong version of Ekman's theory. Within affective science, however, basic emotion theory has evolved. Cowan and Keltner argue that naturalistic expressions are better characterized by 28 discrete categories including amusement, awe, and contemplation [48]. Others relax the tight link between expression and other components. Ekman himself emphasized that expressions may be regulated via situation- or culture-specific display rules [49]. More recent scholars emphasize that the connection between internal emotion and external behavior is mediated by latent variables, shifting from a view of emotion as a fixed circuit (e.g., a Markov-decision process) to a more flexible action program that accounts for context (e.g., a partially-observable Markov-decision process) [50], [51].

Whereas basic emotion theory argues expressions reveal an underlying emotional state, others argue they are deliberative communicative acts designed to shape interaction partners. For example, Fridlund argues that evolution should extinguish any mechanism that gives others unfettered access to our feelings and intentions, as this allows exploitation [52]. Rather, expressions evolved as communicative tools to shape social interactions. If someone displays anger during a conversation, this does not provide evidence about their true emotions, but rather should be seen as akin to saying "back off or I will attack you." In low-stakes or cooperative contexts, whether an expression reflects an underlying emotional state or serves as a communicative act is likely irrelevant. If I'm happy to see a friend, whether my smile reveals true underlying joy (as argued by Ekman) or communicates the intention to affiliate (as argued by Fridlund), the social consequence is probably the same. But in high-stakes situations, this distinction is often seen as important. For example, when someone is experiencing rage, they are likely less able to regulate their expressions and behavior: in other words, the various components of emotion as outlined by Scherer are more likely to be in alignment. More generally, the impact of expressions on observers likely depends on their *perceived authenticity*: i.e., the extent to which observers believe they can make reliable inferences from emotion displays [53].

What do emotional expressions *do*? One benefit of these debates has been to highlight that emotional expressions do far more than simply express emotion, but also impact the behavior of those that observe these expressions. For example, Pelachaud and Poggi [54] and Scarantino [55] build on speech act theory [56] as a framework for characterizing what individuals *do* when they produce an emotional expression during a social interaction. Analogous to the *illocutionary* function of speech, such expressions can reveal the emotions of the sender, but also serve to communicate beliefs, intentions and social requests. Analogous to the *perlocutionary* function of speech, expressions can impact the feelings, thoughts or actions of the audience. And just as with speech, a single expression can perform all of these functions simultaneously. For example, if a listener reacts to an offer in a negotiation with a frown *and* the presence of a frown increases the likelihood that a listener will reject the offer, this expression serves an illocutionary function (i.e., provides probabilistic information about future intentions). If, upon seeing the listener's frown, speakers tend to proactively withdraw or soften their offer, this expression serves a perlocutionary function (i.e., tends to induce concession-making).

Several lines of emotion research have examined how emotional expressions impact the audience (i.e., the perlocutionary function of emotion), and these are perhaps the most relevant to human-machine interaction. Such research can be roughly divided, on the one hand, into theories that treat expressions as revealing *information* about the sender and, on the other hand, into theories that see expressions as *automatically evoking* emotions in the audience even without their conscious awareness (see [27]). This distinction can also have ethical implications as discussed in Section V: i.e., if someone is influenced by an expression without awareness, the expression could be seen as coercive.

Some of our work illustrates the information view. We found that people form expectations about if their partner will collaborate or compete from their partner's facial expressions, and this expectation shapes decisions via a process we call *reverse appraisal* [5], [57]. The intuition here is that if emotions are associated with an appraisal of how an external event impacts a person's goal, then their emotional expression reveals this goal. For example, if a partner smiles after mutual-cooperation, this logically entails they value cooperative acts and will likely cooperate in the future. In contrast, if a partner smiles after exploiting you, this implies they are purely self-interested and will likely exploit you in the future. In line with such "emotion-as-social-information" theories, machines might use expressions to efficiently communicate their mental state to interaction partners in a manner analogous to speech, and as highlighted in Section III, appraisal theories can inform how these expressions should be generated.

In contrast, theories of emotional contagion argue that the expressions of one person directly evoke emotions in the audience [58], [59]. For example, if a negotiator shows anger during a negotiation it might evoke fear in their partner and yield greater concessions [27]. Contagion can be a particularly effective influence tactic as many interpersonal decisions are often driven more by emotion than by rational decision-making. For example, decisions to trust another party are strongly influenced by evoked emotions, particularly when the party is unaware of the source of their feelings [60]. Thus, as occurs between people [61], a machine that smiles might engender trust and thereby enhance cooperation.

B. Human-Machine Interaction Theory & Evidence

Are these social effects of emotion in human-human interaction relevant to human-machine interaction? In the 1990s, several experimental studies began producing evidence that machines could be treated in a social manner [62]–[66]. Nass and colleagues, subsequently, advanced a general theory of human-machine interaction which argued that people will intuitively treat machines in a social manner when engaging with them in social interaction [11]. Some of their experiments indicated that people were polite to machines [65], applied gender stereotypes to machines based on the topic of discussion [66], and formed more positive impressions of machines that were perceived to belong to the same team [63]. In a strict sense, the theory argued that any finding from social psychology would carry to human-machine interaction because humans, as a cognitive heuristic, carried their knowledge from human interaction to their interaction with machines.

Blascovich and colleagues [12] proposed a more refined view, arguing that machines do not always socially influence people but are more likely to do so, the higher the perceived human-likeness of the machine, both in terms of visual appearance and behavior. In line with this view, studies showed that: machines that displayed emotion tended to outperform those that did not when engaging in decision tasks with humans [5], [67]; mirroring human nonverbal behavior in machines increased rapport [26], [68]; and, avatars that looked like the human user were more likely to increase compliance in behavioral change therapy [69].

Evidence from neuroeconomics behavioral studies provided further evidence of important differences – in terms of brain activation and behavior – when people engaged with machines, when compared to humans in social decision making [70]–[75]. In some studies, people tended to show lower activation of the medial prefrontal cortex, a region of the brain implicated in mentalizing (i.e., inferring of others' mental states), when engaging with machines in comparison to humans [70]–[74]. In other studies, people showed lower activation of regions associated with the experience of emotion when engaging with machines than with humans [75], [76]. These findings are broadly in line with evidence from the mind perception literature indicating that, by default, people expect machines to have lower cognitive and affective mental ability when compared to adult humans [77], [78]. Accordingly, empirical studies showed that participants consistently made more favorable decisions to humans than machines in a variety of decision tasks, and showed less guilt when exploiting machines [79]. Implicit in this work, thus, is the notion that appropriate simulation of affective-cognitive ability in machines could enhance human-machine interaction.

Early experiments with emotionally expressive machines focused on how emotional expressions could improve task performance and subjective impressions in education, behavior change, collaboration, games, commerce, and others [29]. These experiments tended to compare machines with and without emotion, often reporting increased perception of empathy but failing to consistently improve task performance [80]–[86]. Various reasons explain these mixed results, including experimental designs where emotion was redundant with other modalities, did not communicate task-relevant information, and ignored the influence of context in shaping interpretation of expressions. Studies indicating that people preferred consistent to inconsistent displays of emotion (e.g., positive facial expressions paired with negative verbal statements) [87], [88] began producing clearer evidence that the effects of emotional expressions could not be explained by the mere presence of certain displays, but what they meant. This message, though, was emphasized by experiments comparing behavioral consequences in decision making tasks, when participants faced expressive machines that showed the exact same displays, but in different contexts [5], [35], [36]. For instance, in a social dilemma, expressing joy following cooperation led to considerably stronger expectations of future cooperation than when the exact same expression was shown following exploitation by the machine. Other studies have since reported further behavioral consequences in various decision making settings [57]. In sum, there is now a growing body of evidence on the subjective, relational, and behavioral effects of emotional expressions in machines, which replicates and extends findings from the human-human interaction literature.

III. Modeling the Derivation of Emotion and Expressive Behavior

Emotional expressions tend to maintain a certain level of coherence across situations. People show pleasure when their goals are satisfied and displeasure when their goals are thwarted. They direct their pleasure and displeasure at the causal agent that satisfied or thwarted their goals. And observers use these patterns, for example using reverse appraisal, to understand and coordinate their responses. If machines wish to leverage the social effects of emotional expressions, maintaining coherence across the signals is important. Computational models of emotion, and appraisal models in particular, are an important tool for maintaining this coherence by ensuring consistency with the agent's beliefs, desires and intentions.

What a designer intends to achieve by consistent emotional behavior is dependent on the function the behavior is realizing, as noted in Section II-A. The function may be to convey the agent's *underlying emotional states*, or what the agent is feeling. It may also, alternatively or jointly, be a *deliberative intentional act* that seeks to influence the observer in some way. Of course, an artificial agent is a designed artifact. From the designer's perspective, the expression is serving the application's intended social function whether the expression is reveling emotional states or deliberative.

Thus, there may be different degrees of coherence depending on whether expressions are intended to be perceived as deliberative communicative acts or meant to portray underlying emotional state. In both cases, a model can ensure the expressions convey an internal state consistent with the agent's beliefs, desires and intentions. An embodied agent modeling a negotiator might express anger when an opponent makes an unfair offer, for example. However, if the expressions are intended to convey the underlying emotional state, we may expect that the emotion also transforms the wide range of

cognitive states and processes in the agent as it copes, much as emotions do in people. Our angry negotiator, for example, may *irrationally* shut down any effort to negotiate and storm out of the room, even if a failed negotiation is a worse outcome for it. A model therefore may seek to achieve a consistency between an agent's various perceptual, reasoning and behavioral processes even as emotion transforms those processes. Such more detailed modeling is especially relevant in high stakes training applications such as using embodied agents in conflict negotiation or as virtual patients in medical communication training in life threatening scenarios.

In keeping with this concern of coherent, consistent expressive behavior, we assume in this section that the derivation of emotion and its expressive consequences is through a computational process model. If the goal is to reflect true emotional state, the model encompasses an agent's assessment of its environment with respect to its goals or concerns. The assessment in turn leads to affective states that influence behavior that explicitly or implicitly express affective information. If the goal is to deliberately communicate emotion, the model can be used to simulate the intended assessment of the expressions by the observers.

The use of a process model to derive an agent's emotional state has certain benefits but also makes assumptions and/or trade-offs. In addition to coherent, consistent expressions, a model may support a generative, flexible capacity to assess a range of events relative to how they impact the agent's goals and then respond in different ways. Such generative flexibility depends in part on how the model characterizes situations and undertakes the assessment. For example, models that realize a hard wired relation between situations and emotions may have a fixed set of situations that they can assess and further won't be responsive to the agent's evolving goals. See the related discussion of hard-wired models [89] as well as the argument for domain-independent modeling of emotion [90]. However, this generative flexibility implies the designer of an agent cedes direct control over the agent's behavior to the model. This may make it harder to circumscribe behavior, which can be of special concern in applications such as mental health and elderly care where the interaction may have significant consequences. Additionally, the need for such flexibility depends on the range of situations the agent faces, the complexity of the agent's concerns and the range of responses it has. In limited scenarios, the modeling can be an exercise in over-engineering. Nevertheless, in large or more open-ended scenarios, it can be critical to have a model to achieve behavior that is consistent with the agent's concerns.

A. From Theory to Model

As discussed in Section II-A, the various theories characterize emotions and emotion processes differently. They focus on different elements of emotion, have different levels of detail of underlying eliciting conditions of emotion and also differ to the degree that they seek to describe the dynamics of emotion. Thus, when leveraging theory to inform a model of emotion, we need to consider how the theory influences design goals for expressive behavior. If the goal is, for example, simply to use behavior to convey a sense of agitation through manipulations of the force and extent of an agent's gestures, a model based on a dimensional theory may suffice. If, instead, we want behavior to convey the target or cause of emotion, such as who is to blame for the agent's anger, we may want an appraisal-based model, where specific appraisals of the desirability of an event and its causal attribution identify whether someone is blameworthy or praiseworthy. We may also want to use theories that go beyond emotion to explore coping strategies, such as the appraisal and coping work of Lazarus and Folkmann [91], [92] where emotion seeks to alter the person's relation to emotion invoking situations by influencing various cognitive states, including attention, beliefs, desires and intentions as well as the processes that maintain those states. Modeling coping can also provide an additional mechanism to deliberately achieve social functions as detailed below.

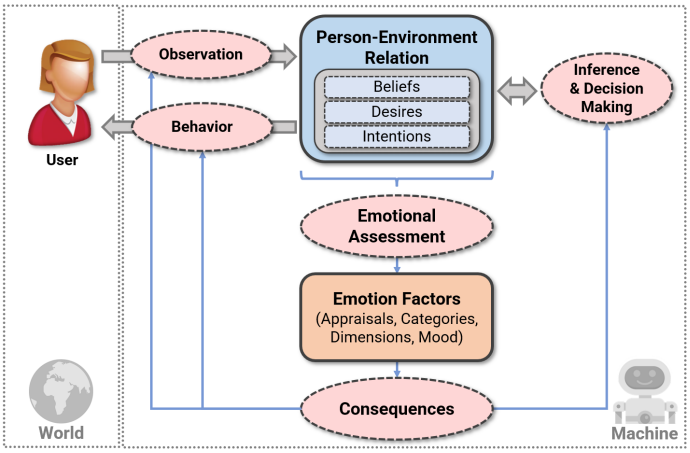

Fig. 2. General framework for computational models of emotion.

B. Modeling of Emotion Causes and Consequences

The construction of a computational model will need to make assumptions about the underlying emotional and cognitive processes which in turn influence expressive behavior. Figure 2 provides a simplified, general characterization of an emotional agent. It depicts the relation of various states or inferences (represented in boxes) and the processes (ovals) that derive those states or inferences. The upper part of the diagram depicts the various perceptual, behavioral and decision-making processes that maintain an agent's subjective representation of its relation to environment, the Person-Environment Relation. This relation represents the agent's beliefs about the world, a characterization of how these beliefs impact the agent's desires, or goals, and its intended actions or plans. This overall framework is typically integrated with decision making processes, such as classical planning, logic-based reasoning, decision-theoretic reasoning, Markov decision processes or reinforcement learning [89]. Beneath these cognitive processes/states is some form of Emotion Assessment process that derives emotional state information. Depending on the model, these can include emotional categories, dimensions, appraisals and/or mood. As depicted by

the left consequence links, the emotional state can influence perception processes and behavior. For example, this may include expressive behaviors that reveal underlying emotional states. The right consequence link represents the influence of emotion on various, more deliberative, cognitive processes that impact beliefs, desires and intentions, as well as indirectly the behaviors. In particular, coping strategies such as avoidance, wishful thinking, distancing and resignation can be modeled as changes in attention, beliefs, desirability and intentions, respectively [93]. This more deliberative process can also model the social function of emotional expression, under the assumption the agent is modeling how that function influences others. For example, an agent feeling fear, which implies a lack of control in a situation, can seek to establish control, cope, by expressing anger.

The details of how these various components are realized impact the nature of the agent's expressive behavior. Here we focus on emotion related components in the lower half of the diagram (see Figure 2). Note most computational works on emotion have relied on some variant of appraisal theory [15], [89] to perform this assessment. Theories that have informed this work on computational models include the work of Ortony et al. [94], Lazarus and Smith [95] and Scherer's Component Process Model [96]. Several factors have led to the dominance of appraisal theories in computational work. Appraisal theories are well suited for modeling since they fit well with Belief-Desire-Intention and decision theoretic models used in works on planning, decision-making and reinforcement learning. Indeed, the assessments of a situation that are critical for the agent's success can be directly tied to appraisal variables. For example, appraisals of goal conduciveness, coping potential/control, novelty can be directly related to decision-making and belief maintenance tasks (e.g., [97]). Further, as argued in [98] it is possible to incorporate other theoretical views of emotion within a generic appraisal framework, such as dimensional theories and emotion categories. Finally, appraisals, by directly relating emotions to an agent's beliefs and desires, provide the mechanisms we seek to ensure coherent, consistent behavior.

Person-Environment Relation: The Person-Environment Relation depicts the agent's subjective inferences about its beliefs about the state of the environment and how it impacts its goals. Depending on the implementation, it can be an explicit representation the agent derives or transient inferences derived during decision-making. This relation may encode how an event facilitates or hinders a goal, whether that event was expected and who/what is responsible for the event. Critically, if the model is going to represent and express emotions about the future, such as hope, or the past, such as regret, then the agent must have a way representing or inferring how past, present or future events may impact its goals. Additionally, this relation may also be from different perspectives, the agent's own perspective but also from the perspective of others. The latter presumes the agent has some model of other agents, for example what their beliefs, desires and intentions may be.

This perspective taking assumes some form of what is typically referred to as theory of mind or mentalizing [99], [100]. This capacity can be used to provide an agent with a range of sophisticated social inferences, such as how others may react to an event. One common approach to modeling theory of mind is to use recursive modeling techniques [101], [102] and appraisal processes have been integrated into such frameworks [103]. Specifically related to theory of mind is the ability to engage in reverse appraisal discussed in Section II-A whereby an agent can draw inferences about what others' beliefs and goals are from its emotional expressions [5], [104]. Maintaining a model of specific others also allows more flexible use of deliberative, communicative behaviors. Instead of an agent negotiator using an expression of anger whenever angry, it can, for example, use a specific model of an observer to infer whether that expression will be effective given this observer and situation. Finally, we should stress that Figure 2 suggests the Person-Environment relation is explicitly represented but that depends on the underlying architecture. These relations can be procedurally inferred on demand as opposed to be explicitly represented.

Assessment: An assessment process derives emotional meaning from the Person-Environment Relation. The level of detail and guidance on how this assessment process works differs markedly across emotion theories. Appraisal theories, however, give explicit guidance about this assessment process through various appraisal variables, which can be directly tied to assessments of a situation that are critical for the agent. For models that employ appraisal, a decision must be made about how appraisals are derived, in particular do they operate in parallel [105] or are there specific constraints on the order in which individual variables are derived [96]. Critically for our current discussion, such dynamics of the model may well be reflected in the dynamics of behaviour. Some emotion theorists have argued for a component based approach to facial expressions whereby individual appraisals drive particular facial actions [98], [106]. For example, an eyebrow lift is the result of appraisals of novelty or unexpectedness. Basic emotion theory [47] argues for canonical mapping from emotion categories to particular patterns of facial expressions. If a designer chooses a component based approach, then the dynamics of the appraisal process impacts facial actions dynamics. Specifically, the dynamics of how the individual appraisals unfold over time impacts the temporal characteristics of the individual facial actions.

Affect Factors: An Affect Derivation process then derives factors that encompass the agent's affective state. These affect factors can include appraisals, emotion categories, dimensions and mood, all of which can influence multimodal behaviors, with differing dynamics. Individual appraisal variables, for example, may be in flux as the person-environment situation is altered either by exogenous events or the agent's own cognitive or behavioral actions. The evaluation of some appraisal variables may take more time to process than others (e.g. novelty check is faster to process than norm compatibility). In contrast, mood can be treated as a lagged variable that changes slowly.

Consequences: These affect factors can have a range of consequences, which may involve direct reflections or unintended leakage of internal mental states. This includes behaviors traditionally associated with emotion, notably facial actions like smiles and furrowed brows, but also physical ac-

tions typically not associated with emotions but which can be executed in a fashion that suggests emotions, such as angrily knocking on a door [107], [108]. Finally, there may be a wide range of more cognitive transformations of beliefs, desires, intentions and actions that may lead an observer to infer underlying emotions. One such transformation can involve exploiting the social function of emotional expressions, such as expressing anger when experiencing fear, as a coping strategy to regain some control. It is also the case, however, that exploiting this social function need not require any underlying emotion state playing a causal role, in which case it becomes just another action in the space of actions designed to influence others' beliefs and behavior.

C. Selecting the Modality of Expressive Behaviors

A fundamental question remains concerning what modalities or forms of expression to use to convey affect and how those modalities are realized in a specific behavior. There are many alternative modalities to consider, including facial actions, posture, head movement, gaze, touch, prosody and dialog. Here we can get guidance again from psychological theory as well as empirical research in both psychology and artificial agents. In the following, we consider several factors.

Type of emotion inference: One key issue to consider is what modalities will readily lead to a particular type of inference in the observer. Research suggests observers make emotion category judgements from facial expressions [47]. Some studies have suggested posture [109], [110] and body motion [107], [108] are useful for conveying more broad affect dimensions such as arousal, though evidence for categorical inferences from static postures and bodily movement has been found as well [111]–[113]. Other studies have looked at gaze posture [114]. Other modalities are discussed in Section IV.

Reliability of Inference: In selecting a modality, we can also consider the reliability of inferences, both from perspectives of the internal mental state of the agent expressing the behavior or, alternatively, the designers' intent concerning the inference they want the observer to make. As noted in Section II-A, the inference from facial action, for example, is not necessarily a correct assessment of the internal state of the agent expressing the emotion. Studies have even shown that people's inferences about the meaning of stereotypical expressions don't necessarily agree, independently of the question of whether it reveals a specific underlying emotional state [115]. Nevertheless, if the goal is to ensure an observer makes a specific inference, it may help ensure greater uniformity across observers by relying on stereotypical expressions or better yet hyper-realism or super-normal behaviors [116], [117] that exaggerate the behavior. Again, this depends on the application. For example, in an application designed to train doctors' communication skills [118], we may want realism, as opposed to a tutoring application used to foster engagement in students [119] where we may want more uniform positive emotional responses and are free to consider non-realistic portrayals of emotion.

Modality Speed and Expression Dynamics: Modalities differ in terms of speed. Facial actions can be comparatively rapid and therefore well suited to convey momentary information. In particular, micro-expressions which are often associated with unintended leakage of internal emotion states [120], can be on the order of tens to hundreds of milliseconds. In contrast, gestures can take seconds. Postural shifts are often both slow and infrequent and, therefore, may be well suited for affect information such as mood that changes slowly and less frequently. However, this will depend on the nature of the event. A highly arousing threat to personal safety is likely to lead to comparatively rapid movement of even large masses resulting in a signal that is hard to misinterpret. Another key factor concerns the dynamics of expressions. In particular, the onset, duration and offset time as well as the symmetry of facial actions influence the inferences people draw [6], [121].

Observer Awareness: Finally, when we consider alternative modalities of expression we need to take into account what is apparent to the observer. As noted, for example, micro-expressions are often small in scale and very brief [122] and therefore hard to perceive. Indeed, people need to be trained to perceive them [123]. So, even though they are reliable signals of leakage, they may not be good choice for conveying leakage to a person interacting with an agent. In essence, there is a tradeoff in terms of realism versus observer awareness. Resolving that tradeoff will depend on the application's goals.

IV. Multimodal Behavior Generation of Emotional Expressions

The previous section discussed the computational processes involved in deriving emotions from the perspective of driving the multimodal behaviors of an agent. In this section, we focus on presenting computational approaches for implementing these multimodal behaviors in machines (Figure 3). Given its prevalence, we focus more on anthropomorphic expression, being more succinct on affective dialog models and prosody. For a more detailed review, especially on the last topics, we refer readers to [124].

Since the first facial animation models, scholars have been interested in modeling expressions of emotions. Many approaches have been proposed. An aim shared by scholars is to endow agents with a large spectrum of emotional expressions that are recognized as such by human participants. Thus, models that generate emotional expressions are validated through perceptive studies to measure the recognition rate of these expressions. The impact of these expressions on human user's performance and stance that arise during an interaction with an agent is also carefully studied.

At first, computational models for the emotional expression of agents focused on facial expression. Body posture and expressivity were added later, and recently touch is also being incorporated. Several models were proposed relying on different theoretical models, namely categorical, dimensional and appraisal models. Others got inspiration from dance studies, observational studies, and perceptual ones.

Early models: Early models of facial expressions in embodied virtual agents [130] were based on the specification of the expressions of the six prototypical emotions [47]. Each expression was defined as a set of Action Units of FACS, a commonly used system to describe facial expressions [131].

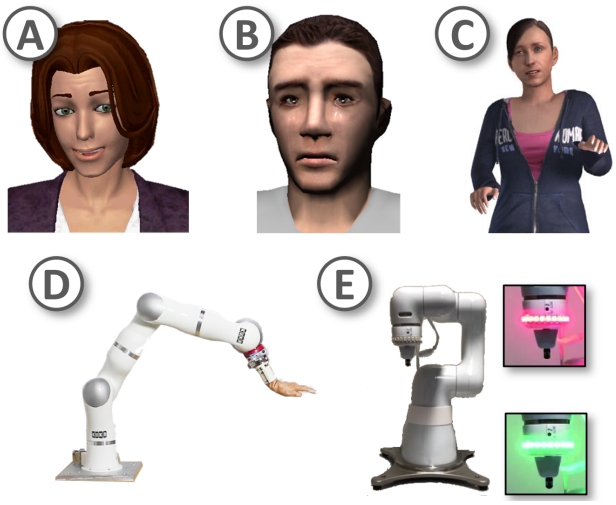

Fig. 3. Social functions of emotion can be expressed in machines in various forms: A, Facial expressions can convey (simple and mixed) emotions [125]; B, Perception of emotion can be enhanced by rendering tears, wrinkles, and blushing [126]; C, Bodily expression combined with facial expression can improve emotion recognition [127]; D, Touch can convey different degrees of valence and arousal [128]; E, Robotic manipulators can be augmented with colored LEDs to communicate emotion [42], [129].

To enlarge the number of emotions that could be displayed on an agent's face, models relied on dimensional representations of emotion [132]. An expression corresponding to an emotion in the 2D or 3D emotion space was computed as a linear interpolation of pre-defined expressions (e.g., the prototypical ones) [133]. These earliest models viewed the expression of an emotion as a static expression at its apex; they did not consider the temporal evolution of an expression. Later models simulated the dynamic display of behavior signals [134], [135]. They implemented the emotional expressions as a sequence of signals emerging from the evaluation of an event as stipulated in appraisal approaches. The Componential Process Model proposed by Scherer [96] views that facial expressions reflect ongoing appraisal of events. Depending on the evaluation of appraisal variables such as novelty, goal obstruction, intrinsic pleasantness, specific facial signals were activated. The computational models built on this approach [134], [135] rendered the expression of emotion as a sequence of signals that were temporally composed on the face. Other approaches relied on observational studies where multimodal signals and their temporal alignment were annotated [136].

Complex expression: An expression may not convey solely one emotion. It could result from the blend of emotional expressions that correspond to the superposition of emotions by, for instance, masking one expression by another (Figure 3-A). To simulate such complexity, the face can be divided into regions. Facial signals on these regions are combined by applying fuzzy rules [125], derived from observations [125] and literature [47].

Skin rendering: Most models presented so far focused on defining the muscular activity on the face and the dynamics of movement. But the emotional expressions may change also the rendering of the face with the appearance of tears or wrinkles, or the change of skin color [126] (Figure 3-B). With the advance of rendering techniques, highly realistic results are obtained, even in real time [137], [138]. This is often seen in video game characters, which have highly realistic appearance in terms of geometry, skin texture, lighting, hair, and clothing. Other techniques make use of databases of images of emotional expression of people. For example, the Expression Generative Adversarial Network (ExprGAN) [139] is an interface to edit photo-realistic facial expressions of humans. ExprGAN uses an encoder-decoder network trained on a small dataset and can generate blends of expressions with different intensities.

Posture and expressivity: Posture and body movement can convey emotions [140] (Figure 3-C). Dance annotations, in particular the Laban annotation schema, offer a set of categories to characterize behavior expressivity. The Emote [141] system implemented the effort and shape categories to modulate the quality of movement. A movement can be executed with more or less force, speed, smoothness, etc. Hartmann et al [142] proposed a model based on the six behavior expressivity parameters resulting from perceptual studies [143], [144], namely the spatial volume of gestures, its velocity, its energy, its repetition, the fluidity between gestures, and their frequency of occurrence. These parameters modify directly the animation of expressions and are applied to the stroke of a behavior. Neff and Fiume [145] proposed animation controllers that act directly on the body joints, obtaining a more precise control to manipulate movement. Breathing dynamics – e.g., fast and deep breathing in anger – have also been shown to be successful in conveying emotion [146].

Touch: While most existing approaches concentrated on the visual modality, face and body, latest studies began exploring touch (Figure 3-D). Social touch can convey a great variety of communicative functions of emotions, as often noted by social roboticists. This type of touch, be a tap or a caress, its dynamic properties such as velocity, pressure, and force can be associated with different emotions [147]. The mapping from touch to emotion is not unique, i.e., a touch can be used to express different emotions and an emotion can be conveyed by different touches. The interaction context helps disambiguate how it is interpreted. To endow agents with touch capabilities [148], two main types of haptic technologies have been envisioned. One allows agents to feel touch by humans. Artificial skin, textile, voice coils, etc., have been used to recreate touch sensations and its dynamic properties [128]. The other, which is pertinent to virtual agents, aims to give humans the impression of touching the agent [149]. Work has also been conducted to create the illusion of touch by manipulating other senses, such as sound [150].

Affective natural language generation: In addition to visual and haptic expressions, verbal behavior can also convey emotion. Initial models selected different vocabularies to convey positive or negative affect [151]. This method is simple but does not produce a great variety of emotions. Moreover, the model acts essentially on the choice of adjectives, which is limited. Latest models rely on deep learning approaches. They take advantage of interesting results obtained using generative adversarial networks (GANs) for natural language generation. Some approaches propose to train their models

on data labeled using sentiment analysis (positive, negative, neutral), text length, and topic [152]. To enlarge the number of possible emotions when generating language, Zhang et al. [153] proposed a cross-domain text sentiment transfer model based on a GAN. The proposed architecture embeds an emotional text generator, a sentiment discriminator, and a domain discriminator. The discriminators are used to pilot the emotional text generation during the training phase and learn emotional patterns.

Voice synthesizer: Whereas the aforementioned methods focused on generating emotional text, several attempts have been carried out to create emotional voice with synthesizers. MaryTT is an open-source speech synthesizer [154] that supports EmotionML, a W3C markup language standard to represent affective states using categories, dimensions, appraisals, and action tendencies [155]. SSML is another markup language whose tags can be inserted in text to indicate how prosody features should be rendered to create expressive voices [156]. Speech rate, fundamental frequency, voice loudness are factors that can be controlled. Earlier voice synthesizers relied on unit selection and made use of specific expressive voice databases that were costly and time consuming to build. Neural text-to-speech synthesizers overcome these limits and offer more natural and expressive synthetic voices [157], [158]. However, the challenge of controlling the voice parameters arises. Zhu and Xue [159] propose to control the emotions strength by training on three datasets containing emotional speech either with weak, medium or strong intensity. Their approach offers control over strength level.

Non-anthropomorphic expression: So far in this section, we have focused on human-like expressions, but it is also possible to achieve the social functions of emotions through non-anthropomorphic expression. Animators have long understood that exaggeration of expressions (e.g., stretching faces) can help audiences connect with characters and suspend their disbelief [160]. In the cinematic, performance, design, and media arts, color and lighting are often used to convey affect to viewers [161]. Through time, music has been used to shape the listeners' emotion and mood [162]. These forms of expression can be readily applied to virtual agents, where the designer has more control over the virtual world (e.g., it is easier to simulate exaggerated facial expressions or change the environment's lighting). In robotic settings, though, there also has been growing interest in exploring some of these forms of expression, especially when it is not possible, practical, or desirable [163] to have complex anthropomorphic designs [42], [129] (Figure 3-E).

Validation studies: These different forms of emotional expressions have been validated through perceptual studies, mainly through recognition tasks [136]. Participants are asked to attach a label chosen among a closed list or describe the expression freely. As it is quite common in this type of study, the recognition task is done without any information of context (e.g., what triggered the emotion and who or what is the target of the expression). The participants just rate the (static or dynamic) expression of the agent. As reviewed in section II-B, further studies can then be conducted where emotional expressions are triggered in human-agent interaction to understand, through objective and subjective measures, the social effects of expressions.

V. Ethics

Just as within interactions between people, the emotional expressions of machines *can* shape the beliefs, actions and emotions of people that interact with these machines. But *should* such techniques be incorporated into applications that impact people's lives and access to opportunities?

Simply the idea of emotional machines seems to evoke strong emotions and extreme ethical positions. AI is commonly seen as rational and emotionless [77] and adding emotion into machines can make them seem disturbing or uncanny [164]. Perhaps because of this, some ethical proposals concerning emotional machines seem disproportionate to other ethical concerns in AI. For example, AI Now recently called for a total ban on the use of affect recognition [165]. Attitudes towards affect generation have been similarly extreme. Several have argued that allowing machines to express emotion is "morally deplorable" as machines don't truly feel [166], apparently oblivious to extensive research and common wisdom that expressions serve important social functions independent of their connection to true feelings (see [167], [168] for similar extreme views). At the other extreme, Scheutz argues that agents not capable of affective communication will inevitably cause humans harm, implying the use of expressive communication is a moral imperative [169].

More nuanced perspectives emphasize that emotional machines raise similar ethical concerns as other AI technology, and thus can be examined within existing ethical frameworks. For example, Cowie reviews a set of key ethical principles and highlights how they can guide the generation of affective systems [170]. These include the principle of *beneficence* (i.e., machines are obliged to act for the benefits of others and balance benefits against potential risks), and *respect for autonomy* (i.e., people interacting with affective systems should be free from factors that subvert their ability to reflect and decide rationally). Issues specific to emotion can be mapped into these broad principles. For example, the focus on *transparency* in AI systems relates to respect for autonomy as people cannot make rational decisions in collaboration with AI if they fail to understand how the AI works. Similarly, if expressions are indeed deceptive, or if seeing emotional expressions shapes a person's decisions without their conscious awareness [171], a person's autonomy has been undermined.

Within affective computing, much of the ethical debate has centered on emotion recognition, but important concerns (and potential benefits) arise from machines that express emotion. Concerning the principle of beneficence, Cowie [172] and Scheutz [169] have argued that the use of appropriate emotional expressions can actually make machines more ethical by sparing people the distress that would otherwise be caused by interactions with emotionally incompetent systems. For example, many commercial conversational agents say "I'm sorry" without actually implementing the social function of regret (i.e., recognizing that an error was made, and forming an intention to perform better in the future). Concerning the

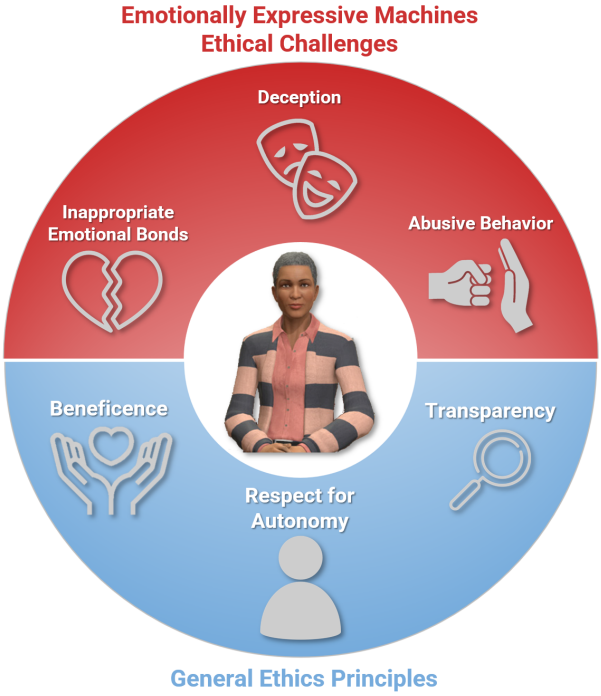

Fig. 4. Emotionally expressive machines should be designed to follow general ethics principles (beneficence, respect for autonomy, and transparency) and tackle specific novel challenges pertaining to deception, inappropriate emotional bonds, and abusive behavior.

principle of respect for autonomy, emotions often function to enhance the autonomy of interaction partners. For example, politeness theory [173] highlights how emotional expressions can help to soften or qualify certain statements in a way that helps buffer threats to a person's "face", including their interactions with technology [3], [174].

Of course, expressive agents can also create unnecessary risks or undermine user autonomy. For example, people can form bonds with machines and will act to protect machines that seem to show fear or pain [175]. This can be problematic if such bonds are not reciprocated [176] or could lead to actual harm if, for example, a soldier's bond with their bomb disposal robot leads them to risk their own lives to save the "life" of the machine. Another issue is the potential for abusive behavior towards expressive machines to transfer to interaction with real people [177]. A similar debate focused on whether violent behavior in video games would transfer to maladaptive action in real life. The simulation of emotional expression in machines may exacerbate the problem by opening the door to new types of abuse – e.g., emotional bullying seen in social networks; a theme that, incidentally, is increasingly being explored by the entertainment industry, as reflected in recent TV shows and movies. Ideally, expressive machines will enable the transfer of *positive* behavior in human-machine interaction to human-human interaction [178], while avoiding the perils of negative behavior. Effective solutions to this challenge likely require a mixture of external factors (e.g., appropriate legislation regulating behavior with machines) and internal factors (e.g., appropriate responses from machines when subjected to abusive behavior [179]).

Just like any other powerful novel technology, there are "good" and "bad" uses for expressive machines. Whereas some may advocate emotional expressions should play an important role in promoting cooperation and the collective welfare, the same technology could be exploited for competition and self-centered goals (e.g., maximization of profit). At the core of the issue is that machines can trivially simulate any emotional expression, independently of their true intentions. In one study, participants engaged in a task with machines that were acting very selfishly, albeit showing emotions that reflected a cooperative orientation (e.g., regret following exploitation) [36]. Interestingly, people concluded that this machine was more likely to cooperate in the future than another which, despite behaving just as selfishly, expressed no emotion. Moreover, whereas some may be quick to condemn deception in machines, a study revealed that some participants endorsed deceptive behavior in machines acting on their behalf during a negotiation task, especially following poor outcomes in prior rounds [180]. In reality, conflict between individual and collective interests permeates human life [181], [182] and it is, in that sense, unsurprising that similar dilemmas would emerge when engaging with machines that may often be representing other humans' conflicting interests.

Even when moral values are at stake – such as when robots are forced into making decisions that weigh in the physical and mental welfare of some versus others – often we find a conflict between normative behaviors, which tend to favor collective welfare, and self-serving behavior. In autonomous driving, for instance, experimental results show that even though most participants preferred autonomous vehicles to maximize preservation of life (even if it meant sacrificing the driver), often they preferred *their* vehicle to prioritize their own life [183]. For reasons like this, some argue we should simply avoid creating machines, or prevent them from being in situations, that are morally charged [184]. However, rather than sacrifice all the potential that expressive machines have for solving some of today's major global challenges – such as the ageing population and the need to assist in care for the elderly or to serve as support of the lack of sufficient educators, especially in under-developed countries – we argue that a more constructive approach is to engage in cross-disciplinary debate to better understand these moral dilemmas and research potential solutions to these challenges, such as recent findings indicating that by considering the nuances of the situation – e.g., the likelihood of injury to involved human parties in autonomous driving [185] – machines are able to produce more satisfactory solutions in moral dilemmas.

VI. Concluding Remarks and Future Challenges

Emotionally expressive machines can facilitate teaming, trust, rapport, and cooperation with human users across a variety of applications. To consistently achieve these effects, we have argued expression generation must be tied to an understanding of social function of emotional expressions in human relationships, but hard challenges must be addressed to validate this claim and realize this vision.

If indeed, expressions serve important social functions, a key challenge is labeled data corpora that facilitate the learning

of these functions. Most corpora focus solely on how third-party observers categorize expressions divorced from the context in which they were produced. More contextual data can be useful to define the parameters of computational models of emotion, as discussed in Section III, train generative expressive models, as reviewed in Section IV, and test and advance emotion theory. Several promising AI trends may address this challenge. First is the emergence of data sets that include context together with emotional expressions (e.g., [186]). A second is the use of synthetic data to train AI systems [187]. Synthetic data is becoming increasingly realistic, given the proliferation of tools such as game engines, and is cheaper, infinite, labelled, and has the potential to avoid ethical complications (e.g., bias and privacy). Moreover, with progress in cognitive modeling, such as described in Section III, it is becoming plausible to generate diverse realistic social simulations. The third trend pertains to so-called "foundational models" that are increasingly able to learn semantic information from large quantities of multimodal loosely labeled data [188]. Some of these systems have, for instance, shown remarkable understanding of abstract concepts such as the (hierarchical) composition of the world and even a basic understanding of math. Paired with algorithms for incremental continual learning [189], these methods may introduce an opportunity to grasp common sense knowledge about socio-affective phenomena with minimal need for specialized datasets. Finally, researchers have begun exploring hybrid modeling approaches that lower the data requirements by leveraging the benefits of data- and theory-driven computational models [190], [191].

The majority of the work in this still nascent field has focused so far on individual and dyadic aspects of the social functions of emotion in machines. Yet, emotion researchers have rightfully noted the importance of considering social function from higher level perspectives including the *group* and, even more broadly, *culture* [192], [193]. In group life, we see emotions serving essential functions, such as signaling social norm violations [194] and encouraging deviant group members to change their behavior [195]. Perceptions of social group membership can, furthermore, moderate how emotion expression is interpreted [196]. At the cultural level, research seeks to understand how cultural values and institutions shape emotion expression in society [197]. One example pertains to cultural rules about the appropriateness of expressing emotion in social settings [198]. Some initial efforts have been made on group and cultural factors shaping the social function of emotion in machines [199]–[201], yet this area is still mostly under-studied and ripe for further investigation.

Much is also still left to be understood about the cognitive-psychological mechanisms driving human emotional expression, which poses a challenge for designers of emotionally expressive machines. A solution is the symbiotic development of expressive agent technology and human theory. In fact, there is a growing tradition of mutually beneficial interdisciplinary research. Researchers have noted, among others, that: computational models of emotion can be used to test and learn about the details of emotion processes, including its dynamic properties [15], [202]; virtual worlds and agents can provide unique advantages, such as experimental control, over other experimental methods such as human confederates [12], [203]; and, machine learning, in particular deep learning methods, can provide valuable insight on emotion phenomena from large collections of publicly available data [204].

Machine expressions raises important ethical challenges. As noted in Section V, we encourage continuous debate involving all stakeholders – researchers, developers, legislators, users, etc. – and focused research to identify acceptable solutions for these ethical dilemmas, with the purpose of building trust in the general public towards this technology. This is especially relevant once we consider long-term interaction with expressive agents and the broader cultural context, yet mostly understudied topics. We have, moreover, only began scratching the surface in this debate. Just as emotional intelligence is highly valued in social and professional life [205], integrated systems not only capable of expressing, but regulating and recognizing emotion will open the door to novel applications, while simultaneously ushering in a new layer of complexity to the ethics discussion.

The next generation of artificial intelligence systems will be *socially intelligent* and capable of comprehending and shaping the social environment it is immersed in. To achieve this, designers cannot afford to ignore the social function of emotional expression, which pervades human life, and leverage it to build trust, rapport, and collaboration with humans. By replicating this social function in machines in ethical ways, thus, we have an opportunity to increase human-machine cooperation and to build AI that is more likely to be trusted, accepted, and adopted by society.

REFERENCES

[1] M. Morris and D. Keltner, "How emotions work: An analysis of the social functions of emotional expression in negotiations," *Res. Org. Behav.*, vol. 22, pp. 1–50, 2000.

[2] N. Frijda and B. Mesquita, "The social roles and functions of emotions," in *Emotion and culture: Empirical studies of mutual influence*, S. Kitayama and H. Markus, Eds. Washington, DC: American Psychological Association, 1994, pp. 51–87.

[3] D. Morand and R. Ocker, "Politeness theory and computer-mediated communication: a sociolinguistic approach to analyzing relational messages," in *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, 2003, p. 10 pp.

[4] R. Niewiadomski and C. Pelachaud, "Affect expression in ecas: Application to politeness displays," *International Journal of Human-Computer Studies*, vol. 68, no. 11, pp. 851–871, 2010.

[5] C. de Melo, P. Carnevale, S. Read, and J. Gratch, "Reading people's minds from emotion expressions in interdependent decision making," *J. Pers. Soc. Psychol.*, vol. 106, pp. 73–88, 2014.

[6] E. Krumhuber, A. Manstead, and A. Kappas, "Facial dynamics as indicators of trustworthiness and cooperative behavior," *Emotion*, vol. 7, pp. 730–735, 2007.

[7] K. Terada and C. Takeuchi, "Emotional expression in simple line drawings of a robot's face leads to higher offers in the ultimatum game," *Front. Psychol.*, vol. 8, 2017.

[8] R. Frank, "Introducing moral emotions into models of rational choice," in *Feelings and emotions*, A. Manstead, N. Frijda, and A. Fischer, Eds. New York, NY: Cambridge University Press, 2004, pp. 422–440.

[9] B. Parkinson and G. Simons, "Affecting others: Social appraisal and emotion contagion in everyday decision making," *Pers. Soc. Psychol. Bull.*, vol. 35, pp. 811–819, 2009.

[10] J. Zaki and C. Williams, "Interpersonal emotion regulation," *Emotion*, vol. 13, pp. 803–810, 2013.

[11] B. Reeves and C. Nass, *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996.

[12] J. Blascovich, J. Loomis, A. Beall, K. Swinth, C. Hoyt, and J. Bailenson, "Immersive virtual environment technology as a methodological tool for social psychology," *Psychol. Inq.*, vol. 13, pp. 103–124, 2002.

[13] C. Breazeal, "Toward sociable robots," *J. Pers. Soc. Psychol.*, vol. 42, pp. 167–175, 2003.

[14] R. Picard, *Affective computing*. MIT Press, 2000.

[15] S. Marsella, J. Gratch, and P. Petta, "Computational models of emotion," in *A blueprint for an affectively competent agent: Cross-fertilization between emotion psychology, affective neuroscience, and affective computing*, K. Scherer, T. Bänziger, and E. Roesch, Eds. Oxford University Press, 2010, pp. 21–45.

[16] J. Bohannon, "Meet your new co-worker," *Science*, vol. 346, pp. 180–181, 2014.

[17] S. Šabanović, C. Bennett, W.-L. Chang, and L. Huber, "Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia," in *Proceedings of the IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, 2013.

[18] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the Autonomous Agents and Multiagent Systems (AAMAS)*, 2014.

[19] G. Gordon, S. Spaulding, J. Westlund, J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[20] W. Swartout, D. Traum, R. Artstein, D. Noren, P. Debevec, K. Bronnenkant, J. Williams, A. Leuski, S. Narayanan, D. Piepol, C. Lane, J. Morie, P. Aggarwal, M. Liewer, J.-Y. Chiang, J. Gerten, S. Chu, and K. White, "Ada and grace: Toward realistic and engaging virtual museum guides," in *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*, 2015.

[21] A. Sloman and M. Croucher, "Why robots will have emotions," in *International Joint Conference on Artificial Intelligence*, Vancouver, Canada, 1981.

[22] G. Bolton, B. Greiner, and A. Ockenfels, "Engineering trust: Reciprocity in the production of reputation information," *Management Science*, vol. 59, no. 2, pp. 265–285, 2013.

[23] H. Simon, *The sciences of the artificial, 3rd edn.* MIT Press, 1996.

[24] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[25] R. Boone and R. Buck, "Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation," *J. Nonverbal Behav.*, vol. 27, pp. 163–182, 2003.

[26] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency, "Virtual rapport," in *Proceedings of the Intelligent Virtual Agents Conference (IVA)*, 2006.

[27] G. van Kleef and S. Côté, "The social effects of emotion," *Annu. Rev. Psychol.*, vol. 73, pp. 1–30, 2022.

[28] J. Cassell, "Embodied conversational agents: Representation and intelligence in user interfaces," *AI Mag.*, vol. 2, pp. 67–83, 2001.

[29] R. Beale and C. Creed, "Affective interaction: How emotional agents affect users," *J. Hum.-Comp. Stud.*, vol. 67, pp. 755–776, 2009.

[30] C. Breazeal and R. Brooks, "Robot emotions: A functional perspective," in *Who needs emotions?: The brain meets the robot*, J. Fellous, Ed. Oxford University Press, 2004.

[31] J. Gratch, J. Rickel, E. Andre, N. Badler, J. Cassell, and E. Petajan, "Creating interactive virtual humans: Some assembly required," *IEEE Intell. Sys.*, vol. 17, pp. 54–63, 2002.

[32] N. Savage, "The slow rise of the caring robot," *Nature*, vol. 601, pp. S8–S10, 2018.

[33] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, "Improving social skills in children with asd using a long-term, in-home social robot," *Sci. Rob.*, vol. 3, 2018.

[34] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi, "Pepper learns together with children: Development of an educational application," in *Proceedings of the IEEE-RAS 15th International Conference on Humanoid Robots*, 2015.

[35] C. de Melo and K. Terada, "The social effects of emotion," *Sci. Rep.*, vol. 10, 2020.

[36] C. de Melo, K. Terada, and F. Santos, "Emotion expressions shape human social norms and reputations," *iScience*, vol. 24, 2021.

[37] C. de Melo, P. Carnevale, and J. Gratch, "The effect of expression of anger and happiness in computer agents on negotiations with humans," in *Proceedings of the Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

[38] C. Gao, "Use new alexa emotions and speaking styles to create a more natural and intuitive voice experience," *Amazon. com, Epub*, vol. 26, 2019.

[39] E. J. Coats, R. S. Feldman, and P. Philippot, "The influence of television on children's nonverbal behavior," in *The social context of nonverbal behavior*, P. Philippot, R. S. Feldman, and E. J. Coats, Eds. Paris: Cambridge University Press, 1999, pp. 156–181.

[40] S. Yilmazyildiz, R. Read, T. Belpeame, and W. Verhelst, "Review of semantic-free utterances in social human–robot interaction," *International Journal of Human–Computer Interaction*, vol. 32, no. 1, pp. 63–85, 2016.

[41] S. Lo, "The nonverbal communication functions of emoticons in computer-mediated communication," *Cyberpsychol. Behav.*, vol. 11, pp. 595–597, 2008.

[42] K. Terada, A. Yamauchi, and A. Ito, "Artificial emotion expression for a robot by dynamic color change," in *Proceedings IEEE RO-MAN*, 2012.

[43] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[44] K. Scherer, "What are emotions? and how can they be measured?" *Soc. Sci. Inf.*, vol. 44, pp. 695–729, 2005.

[45] P. Ekman, R. L. W, and W. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, pp. 1208–1210, 1983.

[46] D. Mackie, T. Devos, and E. Smith, "Intergroup emotions: Explaining offensive action tendencies in an intergroup context," *J. Pers. Soc. Psychol.*, vol. 79, p. 602, 2000.

[47] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, pp. 169–200, 1992.

[48] A. Cowen and D. Keltner, "What the face displays: Mapping 28 emotions conveyed by naturalistic expression," *Am. Psychol.*, vol. 75, 2019.

[49] P. Ekman, E. Sorenson, and W. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[50] R. Adolphs and D. Andler, "Investigating emotions as functional states distinct from feelings," *Emot. Rev.*, vol. 10, pp. 191–201, 2018.

[51] A. Scarantino, "Are ledoux's survival circuits basic emotions under a different name?" *Curr. Opin. Behav. Sci.*, vol. 24, pp. 75–82, 2018.

[52] A. Fridlund, "Evolution and facial action in reflex, social motive, and paralanguage," *Biol. Psychol.*, vol. 32, pp. 3–100, 1991.

[53] S. Côté, I. Hideg, and G. A. van Kleef, "The consequences of faking anger in negotiations," *Journal of Experimental Social Psychology*, vol. 49, no. 3, pp. 453–463, 2013.

[54] I. Poggi and C. Pelachaud, "Performative facial expressions in animated faces," in *Embodied conversational agents*, J. Cassell, Ed. Cambridge, MA: MIT Press, 2000, pp. 155–189.

[55] A. Scarantino, "How to do things with emotional expressions: The theory of affective pragmatics," *Psychol. Inq.*, vol. 28, pp. 165–185, 2017.

[56] J. Austin, *How to do things with words*. Oxford University Press, 1962.

[57] C. de Melo and J. Gratch, "Inferring intentions from emotion expressions in social decision making," in *The social nature of emotion expression: What emotions can tell us about the world*, U. Hess and S. Hareli, Eds. Springer International Publishing, 2019.

[58] E. Hatfield, J. Cacioppo, and R. Rapson, *Emotional Contagion*, ser. Studies in Emotion and Social Interaction. Cambridge: Cambridge University Press, 1994.

[59] J. Tsai, E. Bowring, S. Marsella, and M. Tambe, "Empirical evaluation of computational emotional contagion models intelligent virtual agents," ser. Lecture Notes in Computer Science, H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, Eds. Springer Berlin / Heidelberg, 2011, vol. 6895, pp. 384–397.

[60] J. Dunn and M. Schweitzer, "Feeling and believing: The influence of emotion on trust," *J. Pers. Soc. Psychol.*, vol. 88, pp. 736–748, 2005.

[61] J. Scharlemann, C. Eckel, A. Kacelnik, and R. Wilson, "The value of a smile: Game theory with a human face," *J. Econ. Psychol.*, vol. 22, pp. 617–640, 2001.

[62] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *J. Soc. Iss.*, vol. 56, pp. 81–103, 2000.

[63] C. Nass, B. Fogg, and Y. Moon, "Can computers be teammates?" *International Journal of Human-Computer Studies*, vol. 56, pp. 669–678, 1996.

[64] C. Nass, K. Isbister, and E.-J. Lee, "Truth is beauty: Researching embodied conversational agents," in *Embodied conversational agents*, J. Cassell, Ed. Cambridge, MA: MIT Press, 2000, pp. 374–402.

[65] C. Nass, Y. Moon, and P. Carney, "Are people polite to computers? responses to computer-based interviewing systems," *J. App. Psychol.*, vol. 29, pp. 1093–1110, 1996.

[66] C. Nass, Y. Moon, and N. Green, "Are computers gender-neutral? gender stereotypic responses to computers," *J. App. Psychol.*, vol. 27, pp. 864–876, 1997.

[67] C. de Melo, P. Carnevale, and J. Gratch, "The impact of emotion displays in embodied agents on emergence of cooperation with people," *Presence: Teleoperators Virtual Environ. J.*, vol. 20, pp. 449–465, 2012.

[68] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *Proceedings of the Intelligent Virtual Agents Conference (IVA)*, 2007.

[69] J. Fox and J. Bailenson, "Creating interactive virtual humans: Some assembly required," *Media Psychol.*, vol. 12, pp. 1–25, 2009.

[70] H. Gallagher, J. Anthony, A. Roepstorff, and C. Frith, "Imaging the intentional stance in a competitive game," *NeuroImage*, vol. 16, pp. 814–821, 2002.

[71] K. McCabe, D. Houser, L. Ryan, V. Smith, and T. Trouard, "A functional imaging study of cooperation in two-person reciprocal exchange," *Proc. Nat. Acad. Sci.*, vol. 98, pp. 11 832–11 835, 2001.

[72] T. Kircher, I. Blümel, D. Marjoram, T. Lataster, L. Krabbendam, J. Weber, J. van Os, and S. Krach, "Online mentalising investigated with functional mri," *Neurosci. Lett,*, vol. 454, pp. 176–181, 2009.

[73] S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, and T. Kircher, "Can machines think? interaction and perspective taking with robots investigated via fmri," *PLOS ONE*, vol. 3, pp. 1–11, 2008.

[74] J. Rilling, D. Gutman, T. Zeh, G. Pagnoni, G. Berns, and C. Kilts, "A neural basis for social cooperation," *Neuron*, vol. 35, pp. 395–405, 2002.

[75] A. Sanfey, J. Rilling, J. Aronson, L. Nystrom, and J. Cohen, "A neural basis for social cooperation," *Science*, vol. 300, pp. 1755–1758, 2003.

[76] A. R. von der Pütten, F. Schulte, S. Eimler, L. H. S. Sobieraj, S. Maderwald, M. Brand, and N. Krämer, "A neural basis for social cooperation," *Comp. Hum. Behav.*, vol. 33, pp. 201–212, 2014.

[77] H. Gray, K. Gray, and D. Wegner, "Dimensions of mind perception," *Science*, vol. 315, p. 619, 2007.

[78] A. Waytz, K. Gray, N. Epley, and D. Wegner, "Causes and consequences of mind perception," *Trends Cogn. Sci.*, vol. 149, pp. 383–388, 2010.

[79] C. de Melo, S. Marsella, and J. Gratch, "People don't feel guilty about exploiting machines," *ACM Trans. Comp.-Hum. Int.*, vol. 23, 2016.

[80] K. Hone, "Empathic agents to reduce user frustration: The effects of varying agent characteristics," *Interac. Comp.*, vol. 18, pp. 227–245, 2006.

[81] K. Liu and R. Picard, "Embedded empathy in continuous, interactive health assessment," in *Computer–Human Interaction Workshop on Computer-Human Interaction Challenges in Health Assessment*, 2005.

[82] J. Klein, Y. Moon, and R. Picard, "Online mentalising investigated with functional mri," *Interac. Comp.*, vol. 14, pp. 119–140, 2002.

[83] J. Lester, S. Converse, S. Kahler, T. Barlow, B. Stone, and R. Bhogal, "The persona effect: Affective impact of animated pedagogical agents," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2005.

[84] Y. Lim and R. Aylett, "Feel the difference: a guide with attitude!" in *Proceedings of the International Conference on Intelligent Virtual Agents*, 2007.

[85] H. Maldonado, J. Lee, S. Brave, C. Nass, H. Nakajima, R. Yamada, K. Iwamura, and Y. Morishima, "We learn better together: enhancing elearning with emotional characters," in *Computer Supported Collaborative Learning 2005: The Next 10 Years!*, S. D. C. T. Koschmann, T., Ed. Lawrence Erlbaum Associates, 2005, pp. 408–417.

[86] H. Prendinger, S. Mayer, J. Mori, and M. Ishizuka, "Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces," in *Proceedings of Fourth International Working Conference On Intelligent Virtual Agents*, 2003.

[87] D. Berry, L. Butler, and F. D. Rosis, "Affective interaction: How emotional agents affect users," *J. Hum.-Comp. Stud.*, vol. 63, pp. 304–327, 2005.

[88] C. Creed and R. Beale, "Psychological responses to simulated displays of mismatched emotional expressions," *Interac. Comp.*, vol. 20, pp. 225–239, 2008.

[89] J. Broekens, "Emotion," in *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, B. Lugrin, C. Pelachaud, and D. Traum, Eds. Morgan & Claypool, 2021, pp. 349–384.

[90] J. Gratch and S. Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research*, vol. 5, no. 4, pp. 269–306, 2004.

[91] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press, 1991.

[92] R. S. Lazarus and S. Folkman, *Stress, appraisal, and coping*. Springer publishing company, 1984.

[93] S. C. Marsella and J. Gratch, "Ema: A process model of appraisal dynamics," *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, 2009.

[94] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 2022.

[95] C. S. A and R. Lazarus, "Emotion and adaptation," *Handbook of personality: Theory and research*, vol. 21, pp. 609–637, 1990.

[96] K. Scherer, "Appraisal considered as a process of multilevel sequential checking," in *Appraisal Processes in Emotion: Theory, Methods, Research*, K. Scherer, A. Schorr, and T. Johnstone, Eds. Oxford University Press, 2001, pp. 92–119.

[97] M. Si, S. Marsella, and D. Pynadath, "Modeling appraisal in theory of mind reasoning," *Auton. Agent Multi-Agent Syst.*, vol. 20, pp. 14–31, 2010.

[98] K. Scherer, A. Dieckman, M. Unfried, H. Ellgring, and M. Mortillaro, "Investigating appraisal-driven facial expression and inference in emotion communication," *Emotion*, vol. 21, pp. 73–95, 2021.

[99] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[100] A. Whiten and R. Byrne, *Natural theories of mind: Evolution, development and simulation of everyday mindreading*. B. Blackwell Oxford, UK, 1991.

[101] P. J. Gmytrasiewicz and E. H. Durfee, "A rigorous, operational formalization of recursive modeling." in *ICMAS*, 1995, pp. 125–132.

[102] D. V. Pynadath and S. C. Marsella, "Psychsim: Modeling theory of mind with decision-theoretic agents," in *IJCAI*, vol. 5, 2005, pp. 1181–1186.

[103] M. Si, S. C. Marsella, and D. V. Pynadath, "Modeling appraisal in theory of mind reasoning," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 14–31, 2010.

[104] S. Hareli and U. Hess, "What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception," *Cogn. Emotion*, vol. 24, pp. 128–140, 2010.

[105] J. Gratch, "Émile: marshalling passions in training and education," in *Proceedings of the Fourth International Conferenced on Intelligent Agents*, 2000.

[106] C. Smith and H. Scott, "A componential approach to the meaning of facial expressions," in *The psychology of facial expression*, J. A. Russell and J. M. Fernandez-Dols, Eds., 1997, pp. 229–254.

[107] F. Pollick, H. Paterson, A. Bruderlin, and A. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, pp. B51–B61, 2001.

[108] F. Pollick, "The features people use to recognize human movement style," in *International gesture workshop*, 2003, pp. 10–19.

[109] P. Ekman, "Differential communication of affect by head and body cue," *J. Pers. soc. Psychol.*, vol. 2, p. 726, 1965.

[110] P. Ekman and W. Friesen, "Head and body cues in the judgment of emtoion: A reformulation," *Percep. Motor Skills*, vol. 24, pp. 711–724, 1967.

[111] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *J. Nonverbal Behav.*, vol. 28, pp. 117–139, 2004.

[112] A. Atkinson, W. Dittrich, A. Gemmell, and A. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception*, vol. 33, pp. 717–746, 2004.

[113] K. Walters and R. Walk, "Perception of emotion from body posture," *Bull. Psychon. Soc.*, vol. 24, 1986.

[114] B. Lance and S. Marsella, "The expressive gaze model: Using gaze to express emotion," *IEEE Computer Graphics and Applications*, vol. 30, no. 4, pp. 62–73, 2010.

[115] L. Barrett1, R. Adolphs, S. Marsella, A. Martinez, and S. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Interest*, vol. 20, pp. 1–68, 2019.

[116] D. Barrett, *Supernormal stimuli: How primal urges overran their evolutionary purpose*. WW Norton & Company, 2010.

[117] N. Tinbergen and A. C. Perdeck, "On the stimulus situation releasing the begging response in the newly hatched herring gull chick (larus argentatus argentatus pont.)," *Behaviour*, vol. 3, no. 1, pp. 1–39, 1951.

[118] G. Bogaard, E. H. Meijer, A. Vrij, and H. Merckelbach, "Strong, but wrong: Lay people's and police officers' beliefs about verbal and nonverbal cues to deception," *PloS one*, vol. 11, no. 6, p. e0156615, 2016.

[119] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal, "The persona effect: affective impact of animated pedagogical agents," in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, 1997, pp. 359–366.

[120] P. Ekman, "Lie catching and microexpressions," *The philosophy of deception*, vol. 1, no. 2, p. 5, 2009.

[121] M. Rychlowska, R. E. Jack, O. G. Garrod, P. G. Schyns, J. D. Martin, and P. M. Niedenthal, "Functional smiles: Tools for love, sympathy, and war," *Psychological science*, vol. 28, no. 9, pp. 1259–1270, 2017.

[122] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[123] C. M. Hurley, "Do you see what i see? learning to detect micro expressions of emotion," *Motivation and Emotion*, vol. 36, no. 3, pp. 371–381, 2012.

[124] B. Lugrin, C. Pelachaud, and D. Traum, "The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics," 2021.

[125] J.-C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud, "Multimodal complex emotions: Gesture expressivity and blended facial expressions," *International Journal of Humanoid Robotics. Special issue on Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids*, vol. 20, pp. 477–498, 2006.

[126] C. de Melo and J. Gratch, "Expression of emotions using wrinkles, blushing, sweating and tears," in *Proceedings of the Intelligent Virtual Agents (IVA)*, 2009.

[127] C. Ennis, L. Hoyet, A. Egges, and R. McDonnell, "Emotion capture: Emotionally expressive characters for games," in *Proceedings of motion on games*, 2013, pp. 53–60.

[128] M. Teyssier, G. Bailly, C. Pelachaud, and E. Lecolinet, "Conveying emotions through device-initiated touch," *IEEE Trans. Affect. Comp.*, 2020.

[129] K. Usui, K. Terada, and C. de Melo, "The influence of emotional expressions of an industrial robot on human collaborative decision-making," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2022.

[130] J. Ostermann, "Animation of synthetic faces in MPEG-4," in *Computer Animation'98*, 1998, pp. 49–51.

[131] P. Ekman, W. Friesen, and J. Hager, *THE FACIAL ACTION CODING SYSTEM (Second Edition)*. Salt Lake City: Research Nexus eBook.London: Weidenfeld Nicolson (world), 2002.

[132] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Curr. Psychol.: Dev., Learn., Person., Soc.*, vol. 14, pp. 261–292, 1996.

[133] I. Albrecht, M. Schröder, H.-P. J. Haber, and Seidel, "Mixed feelings: expression of non-basic emotions in a muscle-based talking head," *Virtual Reality*, vol. 8, pp. 201–212, 2005.

[134] M. Paleari and C. Lisetti, "Psychologically grounded avatars expressions," in *First Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence*, 2006.

[135] M. Courgeon, C. Céline, and J.-C. Martin, "Modeling Facial Signs of Appraisal During Interaction: Impact on Users' Perception and Behavior," *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, pp. 765–772, 2014.

[136] R. Niewiadomski, S. Hyniewska, and C. Pelachaud, "Constraint-based model for synthesis of multimodal sequential expressions of emotions," *IEEE Trans. Affect. Comput.*, vol. 2, pp. 134–146, 2011.

[137] J. Jimenez, J. E. I, C. Oat, and D. Gutierrez, "Practical and realistic facial wrinkles animation," in *GPU Pro 360: Guide to Geometry Manipulation*, 2018, pp. 95–107.

[138] R. McDonnell and B. Mutlu, "Appearance," *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pp. 105–146, 2021.

[139] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32.

[140] H. Wallbott and K. Scherer, "Cues and channels in emotion recognition," *J. Pers. Soc. Psychol.*, vol. 24, 1986.

[141] D. Chi, M. Costa, L. Zhao, and N. Badler, "The EMOTE model for effort and shape," in *Siggraph 2000, Computer Graphics Proceedings*, 2000, pp. 173–182.

[142] B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud, "Design and evaluation of expressive gesture synthesis for embodied conversational agents," in *Third International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS)*, Utretch, July 2005.

[143] H. Wallbott, "Bodily expression of emotion," *Eur. J. Soc. Psychol.*, vol. 28, pp. 879–896, 1998.

[144] P. Gallaher, "Individual differences in nonverbal behavior: Dimensions of style," *J. Pers. Soc. Psychol.*, vol. 63, pp. 133–145, 1992.

[145] M. Neff and E. Fiume, "Aer: Aesthetic exploration and refinement for expressive character animation," in *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2005, pp. 161–170.

[146] C. de Melo, P. Kenny, and J. Gratch, "Real-time expression of affect through respiration," *Comput. Animat. Virtual Worlds*, vol. 21, 2010.

[147] M. Hertenstein, R. Holmes, M. McCullough, and D. Keltner, "The communication of emotion via touch." *Emotion*, vol. 9, p. 566, 2009.

[148] F. Boucaud, C. Pelachaud, and I. Thouvenin, "Decision model for a virtual agent that can touch and be touched," in *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 2021.

[149] M. Grandidier, F. Boucaud, I. Thouvenin, and C. Pelachaud, "Softly: Simulated empathic touch between an agent and a human," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2795–2797.

[150] M. Auvray and E. Myin, "Perception with compensatory devices: From sensory substitution to sensorimotor extension," *Cogn. Sci.*, vol. 33, pp. 1036–1058, 2009.

[151] M. Rehm, B. Endrass, and M. Wissner, "Integrating the user in the social group dynamics of agents," in *Proceedings of Social Intelligence Design (SID)*, 2007.

[152] J. Chen, Y. Wu, C. Jia, H. Zheng, and G. Huang, "Customizable text generation via conditional text generative adversarial network," *Neurocomputing*, vol. 416, pp. 125–135, 2020.

[153] R. Zhang, Z. Wang, K. Yin, and Z. Huang, "Emotional text generation based on cross-domain sentiment transfer," *IEEE Access*, vol. 7, pp. 100 081–100 089, 2019.

[154] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the mary tts platform," in *12th Annual Conference of the International Speech Communication Association-Interspeech 2011*. ISCA, 2011, pp. 3253–3256.

[155] F. Burkhardt, C. Pelachaud, B. Schuller, and E. Zovato, "Emotionml," in *Multimodal Interaction with W3C Standards*, 2017, pp. 65–80.

[156] P. Taylor and A. Isard, "Ssml: A speech synthesis markup language," *Speech commun.*, vol. 21, pp. 123–133, 1997.

[157] M. Aylett, A. Vinciarelli, and M. Wester, "Speech synthesis for the generation of artificial personality," *IEEE Trans. Affect. Comput.*, vol. 11, pp. 361–372, 2017.

[158] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[159] X. Zhu and L. Xue, "Building a controllable expressive speech synthesis system with multiple emotion strengths," *Cogn. Syst. Res.*, vol. 59, pp. 151–159, 2020.

[160] O. Johnston and F. Thomas, *The illusion of life: Disney animation*. Disney Editions, 1995.

[161] C. de Melo and A. Paiva, "Environment expression: Expressing emotions through cameras, lights and music," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2007.

[162] P. Juslin and J. Sloboda, *Music and emotion: Theory and research*. Oxford University Press, 2001.

[163] M. Mori, "The uncanny valley: The original essay by masahiro mori," *IEEE Spectrum*, 2012.

[164] J.-P. Stein and P. Ohler, "Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting," *Cognition*, vol. 160, pp. 43–50, 2017.

[165] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, and A. N. Sánchez, "Ai now 2019 report," *New York, NY: AI Now Institute*, 2019.

[166] R. Sparrow, "The march of the robot dogs," *Ethics and Information Technology*, vol. 4, pp. 305–318, 2002.

[167] S. Bringsjord and M. Clark, "Red-pill robots only, please," *Affective Computing, IEEE Transactions on*, vol. 3, pp. 394–397, 2012.
[168] M. Coeckelbergh, "Are emotional robots deceptive?" *IEEE Trans. Affect. Comput.*, vol. 3.
[169] M. Scheutz, "The affect dilemma for artificial agents: Should we develop affective artificial agents?" *IEEE Trans. Affect. Comput.*, vol. 3, pp. 424–433, 2012.
[170] R. Cowie, "Ethical issues in affective computing," *The Oxford Handbook of Affective Computing*, p. 334, 2015.
[171] P. Winkielman, K. Berridge, and J. Wilbarger, "Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value," *Pers. Soc. Psychol. Bull.*, vol. 31, pp. 121–135, 2005.
[172] R. Cowie, "The good our field can hope to do, the harm it should avoid," *IEEE Trans. Affect. Comput.*, vol. 3, pp. 410–423, 2012.
[173] P. Brown and S. Levenson, *Politeness: Some Universals in Language Usage*. New York, NY: Cambridge University Press, 1987.
[174] N. Wang and W. Johnson, "The politeness effect in an intelligent foreign language tutoring system," in *Intelligent Tutoring Systems*. Springer, 2008, pp. 270–280.
[175] A. Horstmann, N. Bock, E. Linhuber, J. Szczuka, C. Straßmann, and N. Krämer, "Do a robot's social skills and its objection discourage interactants from switching the robot off?" *PLOS One*, vol. 13, p. e0201581, 2018.
[176] M. Scheutz, "The case for explicit ethical agents," *AI Mag.*, vol. 38, pp. 57–64, 2017.
[177] A. Knott, M. Sagar, and M. Takac, "The ethics of interaction with neurorobotic agents: a case study with babyx," *AI Ethics*, vol. 2, pp. 115–128, 2022.
[178] R. Aylett, M. Vala, P. Sequeira, and A. Paiva, "Fearnot! – an emergent narrative approach to virtual dramas for anti-bullying education," in *Proceedings International Conference on Virtual Storytelling*, 2007.
[179] L. Fassler, "We tested bots like siri and alexa to see who would stand up to sexual harassment," *Quartz*, 2017. [Online]. Available: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/
[180] J. Mell, G. Lucas, S. Mozgai, and J. Gratch, "The effects of experience on deception in human-agent negotiation," *J. Artif. Intell. Res.*, vol. 68, pp. 633–660, 2020.
[181] P. Kollock, "Social dilemmas: The anatomy of cooperation," *Annu. REv. Sociol.*, vol. 24, pp. 183–214, 1998.
[182] D. Rand and M. Nowak, "Human cooperation," *Tr. Cogn. Sci.*, vol. 17, pp. 413–425, 2013.
[183] J. Bonnefon, A. Shariff, and I. Rahwan, "Risk of injury in moral dilemmas with autonomous vehicles," *Science*, vol. 352, pp. 1573–1576, 2016.
[184] A. Sharkey, "Can we program or train robots to be good?" *Ethics Inf. Technol.*, vol. 22, pp. 283–295, 2020.
[185] C. de Melo, S. Marsella, and J. Gratch, "Risk of injury in moral dilemmas with autonomous vehicles," *Front. Robot. AI*, vol. 7, p. 572529, 2016.
[186] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755–2766, 2020.
[187] C. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," *Trends Cogn. Sci.*, vol. 26, pp. 174–187, 2022.
[188] M. Scheutz, "On the opportunities and risks of foundational models," *arXiV*, p. arXiv:2108.07258v2, 2021.
[189] R. Hadsell, D. Rao, A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends Cogn. Sci.*, vol. 24, pp. 1028–1040.
[190] M. Scheutz, "Towards a prediction and data driven computational process model of emotion," *IEEE Trans. Affect. Comput.*, vol. 12, pp. 279–292, 2021.
[191] D. Ong, H. Soh, J. Zaki, and N. Goodman, "Applying probabilistic programming to affective computing," *IEEE Trans. Affect. Comp.*, vol. 12, pp. 306–317, 2019.
[192] D. Keltner and J. Haidt, "Social functions of emotions at four levels of analysis," *Cogn. Emot.*, vol. 13, pp. 505–521, 1999.
[193] J. Spoor and J. Kelly, "The evolutionary significance of affect in groups: Communication and group bonding," *Group Process. Intergroup Relat.*, vol. 7, pp. 398–412, 2004.
[194] S. Hareli, O. Moran-Amir, S. David, and U. Hess, "Emotions as signals of normative conduct," *Cogn. Emotion*, vol. 27, pp. 1395–1404, 2013.
[195] M. Heerdink, G. van Kleef, A. Homan, and A. Fischer, "On the social influence of emotions in groups: Interpersonal effects of anger and happiness on conformity versus deviance," *J. Pers. Soc. Psychol.*, vol. 105, pp. 262–284, 2013.
[196] J. V. der Schalk, A. Fischer, B. Doosje, D. Wigboldus, S. Hawk, M. Rotteveel, and U. Hess, "Convergent and divergent responses to emotional displays of ingroup and outgroup," *Emotion*, vol. 11, pp. 286–298, 2011.
[197] H. Elfenbein, M. Beaupre, M. Levesque, and U. Hess, "Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions," *Emotion*, vol. 7, pp. 131–146, 2007.
[198] E. Ji, L. Son, and M. Kim, "Emotion perception rules abide by cultural display rules," *Exp. Psychol.*, vol. 69, pp. 83–103, 2022.
[199] C. de Melo and K. Terada, "Cooperation with autonomous machines through culture and emotion," *PLOS One*, vol. 14, p. e0224758, 2019.
[200] R. Aylett and A. Paiva, "Computational modelling of culture and affect," *Emotion Rev.*, vol. 4, pp. 253–263, 2011.
[201] B. Lugrin and M. Rehm, "Culture for socially interactive agents," in *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents*, B. Lugrin, C. Pelachaud, and D. Traum, Eds. ACM Books, 2022, pp. 463–494.
[202] T. Wehrle and K. Scherer, "Toward computational modeling of appraisal theories," in *Appraisal Processes in Emotion: Theory, Methods, Research*, K. Scherer, A. Schorr, and T. Johnstone, Eds. Oxford University Press, 2001, pp. 350–365.
[203] C. de Melo, P. Carnevale, and J. Gratch, "Using virtual confederates to research intergroup bias and conflict," in *Best Paper Proceedings of the Annual Meeting of the Academy of Management (AoM)*, 2014.
[204] A. Cowen, D. Keltner, F. Schroff, B. Jou, H. Adam, and G. Prasad, "Sixteen facial expressions occur in similar contexts worldwide," *Nature*, vol. 589, pp. 251–257, 2020.
[205] P. Salovey and J. Mayer, "Emotional intelligence," *Imagin. Cogn. Pers.*, vol. 9, 1990.

Celso de Melo is a Computer Scientist at the DEVCOM US Army Research Laboratory. He completed his Ph.D. in CS at the University of Southern California (USC) in 2012. His cross-disciplinary research focuses on the development of socially intelligent machines, with a particular focus to the interpersonal effects of emotion in decision making. He is an Associate Editor of IEEE's Transactions on Affective Computing.

Jonathan Gratch is a Research Professor of Computer Science and Psychology at USC and USC's Institute for Creative Technologies. He completed his Ph.D. in CS at the University of Illinois (UIUC) in 1995. He explores models of social cognition to give insight into human and human-machine interaction. He is the founding EIC (retired) of IEEE's Transactions on Affective Computing, and Fellow of AAAI, AAAC and CogSci.

Stacy Marsella is a Professor of Computer Science and Psychology at Northeastern University, as well as a Professor in the School of Psychology and Neuroscience at the University of Glasgow. He received his Ph.D. from Rutgers University in 1994. His interest is in computational models of human cognitive, emotional and social behavior, both as a basic research methodology in psychology as well as the positive application of these models.

Catherine Pelachaud is CNRS Director of Research at ISIR, Sorbonne University. She received her Ph.D. in CS from University of Pennsylvania in 1991. Her research interest includes socially interactive agent, nonverbal communication (face, gaze, gesture and touch), and social interaction. With her research team, she has been developing an interactive virtual agent platform, Greta, that can display emotional and communicative behaviors. She is co-editor of the ACM handbook on socially interactive agents.

VII. Acknowledgement

This work was partially funded by the the Army Research Office under Cooperative Agreement Number W911NF-20-2-0053 and ANR-DFG-JST Panorama and ANR-JST-CREST TAPAS (19-JSTS-0001-01) projects.The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.