Affective Robots Need Therapy

PAUL BUCCI, University of British Columbia Computer Science
DAVID MARINO, McGill Centre for Intelligent Machines
IVAN BESCHASTNIKH, University of British Columbia Computer Science

Emotion researchers have begun to converge on the theory that emotions are psychologically and socially constructed. A common assumption in affective robotics is that emotions are categorical brain-body states that can be confidently modeled. But if emotions are constructed, then they are interpretive, ambiguous, and specific to an individual's unique experience. Constructivist views of emotion pose several challenges to affective robotics: first, it calls into question the validity of attempting to obtain objective measures of emotion through rating scales or biometrics. Second, ambiguous subjective data poses a challenge to computational systems that need structured and definite data to operate. How can a constructivist view of emotion be rectified with these challenges?

In this article, we look to psychotherapy for ontological, epistemic, and methodological guidance. These fields (1) already understand emotions to be intrinsically embodied, relative, and metaphorical and (2) have built up substantial knowledge informed by everyday practice. It is our hope that by using interpretive methods inspired by therapeutic approaches, HRI researchers will be able to focus on the practicalities of designing effective embodied emotional interactions.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → **Cognitive science**; • **Applied computing** → **Psychology;**

Additional Key Words and Phrases: Emotion, embodied, affective, constructivism, therapeutic care

ACM Reference format:
Paul Bucci, David Marino, and Ivan Beschastnikh. 2023. Affective Robots Need Therapy. *ACM Trans. Hum.-Robot Interact.* 12, 2, Article 20 (March 2023), 22 pages.
https://doi.org/10.1145/3543514

1 INTRODUCTION

When a dog wags its tail, is it nervous or happy? The answer is likely either or both depending on the context. Did the dog's owner just come home? Is the dog's owner looking upset because they noticed a broken lamp in the living room? Does the dog usually get a treat when the owner comes home? Did the dog have a good day or a bad day? Let's say we wanted to build a dog happiness detection system into a robot. It would be difficult to ascertain ground truth, because dogs cannot tell us how they feel. At best, we could ask people who know the dog pretty well to interpret the

Authors' addresses: P. Bucci and I. Beschastnikh, University of British Columbia Computer Science, 2366 Main Mall 201, Vancouver, BC, Canada, V6T 1Z4; emails: {pbucci, bestchai}@cs.ubc.ca; D. Marino, McGill Centre for Intelligent Machines, 3480 University Street, Montreal, QC, Canada, H3A 0E9; email: dmarino@cim.mcgill.ca.

2573-9522/2023/03-ART20
https://doi.org/10.1145/3543514

dog's behavior and provide labels for a classifier. But, realistically, we would be better off building a dog-tail-wagging detection system because it would be grounded in observable quantities and just leave out the question of happiness.

Because humans have the ability to introspect, rationalize, represent, and report on our subjective experiences, we believe that we can draw strong connections between observable things like facial expressions and unobservable things like the subjective experience of feeling an emotion. However, research has shown that observable phenomena such as emotional gestures, physiological signals, and even brain activations are not consistent between people who report having the same labeled emotion [Clore and Ortony 2013]. This has led emotion researchers to theorize that emotions are better understood as psychologically and socially constructed individual experiences rather than universal, categorical experiences [Barrett and Russell 2014]. Building a human happiness detection system may be more like building a dog happiness detection system than we would like to admit: it may be possible to determine whether a facial expression is a smile, but determining how the person behind the smile feels requires a level of interpretation that is not appropriate for a classifier.

If emotions are constructed, then we can think of an emotional experience not as caused by any singular portion of a robot's body or behavior but as emerging from the interaction as a whole, contingent on an interactor's narrative framing [Bucci et al. 2018, 2019; Marino et al. 2017]. Instead, the emotional meaning of specific behaviors is continually grounded [Jung 2017; Leahu and Sengers 2014] through ongoing behaviors during the interaction. Recognizing that robots need to be programmed with structured, determinable data, we believe that the answer is not simply to give up on quantitative methods but instead to embed them in constructivist philosophical approaches and methodologies. If you are studying objective phenomena, use objective methods. If you are studying subjective phenomena, use subjective methods. If you are studying both, use mixed methods. Objective approaches are useful for studying phenomena that are objective, determinable, repeatable, and somewhat culturally independent. But much of the emotional phenomena we wish to study within the field of affective robotics are not objective, and subjective approaches are more appropriate when interpretation is fundamental to the phenomena at hand.

We propose that affective robotics researchers incorporate methodologies from therapeutic fields into their scientific approach. Practitioners in these fields have years of experience in dealing with the concrete realities of the relative and interpretive nature of emotions, and yet still undertake quantitative measurement as a matter of course. Specifically, in this article, we look at manualized therapeutic approaches (i.e., **cognitive-behavioral therapy (CBT)** and **dialectical-behavioral therapy (DBT)**), along with somatic, narrative, and trauma-informed approaches [Bath 2008]. Therapeutic methods can offer a theoretical approach to studying emotions that is distinct from current popular qualitative methods. This therapeutically inspired approach would be particularly effective in the domain that affective roboticists care to research: real-life experiences of emotion [Risjord 2011].

In this article, we outline our understanding of how embodied affective robotics could benefit from the lessons of therapeutic practices in physical and mental healthcare. To support our proposal that we can learn from therapeutic care to make better robot bodies and behaviors, we outline a framework that relates different types of emotional phenomena to theoretical bases in psychology and social sciences. Rather than taking an approach that purports to have a single theoretical framework for emotional understanding, we articulate the different ontological assumptions of affective robotics and critique them by presenting practices and assumptions from pain management science and psychotherapy. To assuage concerns having to do with theory of science questions (e.g., "how do we know what we know?" or "how do we prove something works?"), we draw analogies between these practical therapeutic fields and the scientific questions we approach in affective robotics.

We present concrete examples of how to incorporate therapeutic ethics and methods into study design, as well as the theoretical motivation for expanded HRI methodologies. We contribute the following:

(1) A synthesis of the theoretical and pragmatic basis of therapeutic care methods and their meaning for affective robotics
(2) An account of the constructed nature of emotions in HRI and errors that can result from not accounting for emotions as constructed in study design
(3) Resolutions to the preceding and accompanying methodological recommendations based in examples from therapeutic methods.

2 WHY EMOTIONS AS CONSTRUCTED MATTERS TO HRI RESEARCHERS

What does it mean when we say that emotions are constructed? There are two related but distinct senses in which we mean that emotions are constructed: psychologically and socially constructed [Barrett 2017]. Psychologically constructed refers to the phenomenon of our emotional experiences being "trained" into our brain over a lifetime, and activated in the moment as a series of interconnected networks of neurons. Socially constructed refers to the phenomenon that our emotional experiences are created (historically and in the moment) while interacting with other humans and the world. Contrast this to the concept that emotions are available to us *a priori* (i.e., that all humans experience anger in exactly the same way). Understanding emotions as constructed means that each person will have very different memories, sensations, and in-the-moment experiences encapsulated in the same emotion word, such as "anger." Different personal experiences mean different brain structures: there would be biophysical differences as your brain is being constructed (trained) through many social interactions, which we can refer to as a "cultural embedding."

Intuitively, we can use an analogy of sports to understand why biophysics can be both culturally embedded and highly personalized: a weight-lifter will have a different body structure depending on their culture and personal preferences. Their body will depend on the people/places they interact with—for example, their personal trainer will prefer certain exercises, the gym will have only certain equipment, and their nutritionist will suggest specific supplements. Similarly, the weight-lifter's friends might value certain body shapes that will influence the weight-lifter's values about what to practice. In addition, on any specific day, the weight-lifter's immediate biophysical structure is contingent on other cultural and personal preference factors such as their breakfast, whether they stayed up late streaming television shows, and so forth.

The experience of an emotion is only available to the subjectivity of the person experiencing it. Yet in HRI, we rely on objective methods of measurement (e.g., sensors) and statistical methods (e.g., surveys) that treat emotions like they are universally experienced. Constructed emotions is a different understanding of the nature of emotions than is common in HRI. To study emotions from this ontological perspective requires a different epistemology. An *epistemic claim* is one about the way in which we come to know something—that is, how we can study and produce knowledge about a phenomenon. In the next section, we articulate three epistemic approaches and their relevance for HRI.

2.1 Epistemology of Modern Science and Errors in HRI

The scientific method is generally thought of as positing hypotheses that are tested in experimental environments where variations between trials can be causally attributed to controlled variables. Modern methodologies, especially when pertaining to social and psychological phenomena, acknowledge the likelihood of experimenter bias and try to account for this with statistical tests, blind coding, and so on. This statistical approach to scientific causal claims is (somewhat confusingly) referred to as *post-positivism*, meaning that we expect a scientist to posit logical claims but to also to demonstrate the statistical bounds of their claims (in contrast to mathematicians who can simply posit claims and need no experimental demonstration) [Yilmaz 2013].

Put simply, modern scientists agree that there is a real, physical world that we are testing, but that the best we can do to understand the world is make probabilistic causal claims within a determined confidence interval, and attempt to manage bias through careful experimental design.

By contrast, the kinds of phenomena that HRI researchers are interested in studying are often difficult to fit into an experimental design. This is because we often study robots that interact with participants. In this context, the robot is presented as a social actor. Although there are appropriate places to use an experimental methodology in HRI research, we claim that study designs of in situ emotions require constructivist epistemologies.

Constructivism honors the fact that the human experience of reality is a subjective experience that is influenced by culture and prior experience as well as physical reality. Constructivist epistemologies imply research methodologies that can help avoid errors made by assuming that everyone's experience of reality is described in the same way. In the following, we describe four such errors that we believe to be important for HRI to consider: categorical, methodological, instrumental, and social complexity.

We use a running example of studying "trust" via **galvanic skin response (GSR)** and provide four errors that are introduced into experimental methodology by avoiding the constructed nature of emotions (the authors themselves have made these errors numerous times). We use this example as a stand-in for HRI studies that take an emotional phenomenon (trust, love, etc.) and purport to provide a causal link between that phenomenon and a signal (GSR, heart rate variability, robot pose, etc.).

Categorical error. Trust is better understood as an emotional construct or concept that includes a variety of contingent emotions rather than an emotion itself [Holth 2001; Simpson 2007]. By studying trust without making this distinction, the researcher makes a categorical error. The reason behind the categorical error is that trust emerges from cross-cutting ontological[1] and epistemic domains. In other words, trust exists as a combination of somatic, behavioral, and cognitive aspects that are embedded in a cultural frame. In other words, our in-the-moment body feelings and senses, action, and thoughts are constructed from a lifetime of experiences with other people. As a result, measuring trust is like measuring weight lifting—you can quantify aspects of weight lifting, but it makes little sense to ask people to rate "weight lifting" on a scale of 1 to 5. A better design would ask participants to inspect the constituent emotions behind trust. One approach to resolving categorical errors is to *ground* [Jung 2017] experimental terminology to ensure common understanding between researchers and participants.

Methodological error [Schwarz 2009]. Trust is experienced in highly individualized ways that are hard to attend to and communicate—that is, the subjective experience of trust-related emotions will include different bodily sensations, memories, and beliefs in different people. Participants are generally not trained in the introspective methods required to notice these different phenomena. Introspective methods take years of training to master; initial subjective reports have been shown to be elevated [Shrout et al. 2018], which indicates that the measurement process itself can influence measurement values.

Expecting participants in a study to introspect on their emotions without training will introduce uncontrolled and hidden variability. One way to account for this methodological error is including

[1] An ontological claim is one about the nature or existence of something.

training into the study design. A sufficiently high sample size can also give insights to population-level trends but also obscure individual experience.

Instrumental error. Trust is communicated via gestures and words with meanings that require grounding—that is, the meaning of a "smile" or emotion words like "happy" can be ambiguous between interlocutors unless common ground is established through interaction [Jung 2017; Nevill and Lane 2007]. If the researcher does not establish common ground by asking what a participant means when they talk about "trust," the study instrument may not be measuring what the researcher expects.

Social complexity error. An experience of trust is a dynamic, chaotic, and complex phenomenon that (1) relies on affective changes moment to moment, (2) is highly sensitive to conditions, (3) occurs via many interconnected internal brain-body systems, and (4) depends on in-the-moment social processes as well as long-term social processes. Many of these are only understandable through interpretive and inferential social scientific processes. If we understand the human cognitive experience to be formally complex, then we may be dealing with an intractable set of hidden variables that require more rigor within qualitative analyses [Byrne and Callaghan 2013].

We understand that it may seem like we are asking scientists to relax experimental standards if we suggest using interpretive methods, but, in fact, we believe it is the opposite. A solid theoretical understanding of emotions as constructed entails more scientific rigor, yet with the difficult task of incorporating subjective methodologies.

3 UNDERSTANDING THE CONSTRUCTED NATURE OF EMOTIONS

If we accept that emotions are constructed, we must also accept that the phenomena we are interested in when we study the subjective experience of robot bodies and behaviors are so highly context sensitive that it requires approaching with relative, interpretive methodologies. Constructivist epistemologies and methodologies provide a basis of understanding what science and knowledge production means for subjective phenomena [Raskin 2002], but we argue that the best source of theoretical and practical guidance is expert practitioners in trauma-informed care fields who deal with the on-the-ground difficulty of applying introspective methods daily. In this section, we (1) discuss the biophysical motivation for understanding emotions to be constructed, (2) present a worked example of constructed emotions, and (3) present evidence from emotions researchers that have led to a constructivist movement in psychology.

3.1 Emotions Happen All Over the Brain and Body

We believe that having a good understanding of how emotion happens in the brain and body can give a working mental model of the different kinds of emotional phenomena we attempt to study when designing robot bodies and behaviors. In particular, we believe that it gives us a good understanding of why emotion experiences may be different between different people—an increasingly common viewpoint among emotion theorists. In fact, Ortony and Clore [2013] (of the OCC cognitive appraisal theory of emotion) present a summary of evidence against conceiving of emotions as universal experiences:

> *Should one assume then that specific emotions do not exist? No, but perhaps some long-standing assumptions about them should be reexamined . . . emotions are not marked by distinctive behaviors or even by reliable patterns of feeling . . . Many assumed that affective neuroscience would rescue the study of emotion from this untidiness. However, a recent meta-analysis of imaging results concludes that the evidence that specific emotions have specific locations in the brain is not strong.*

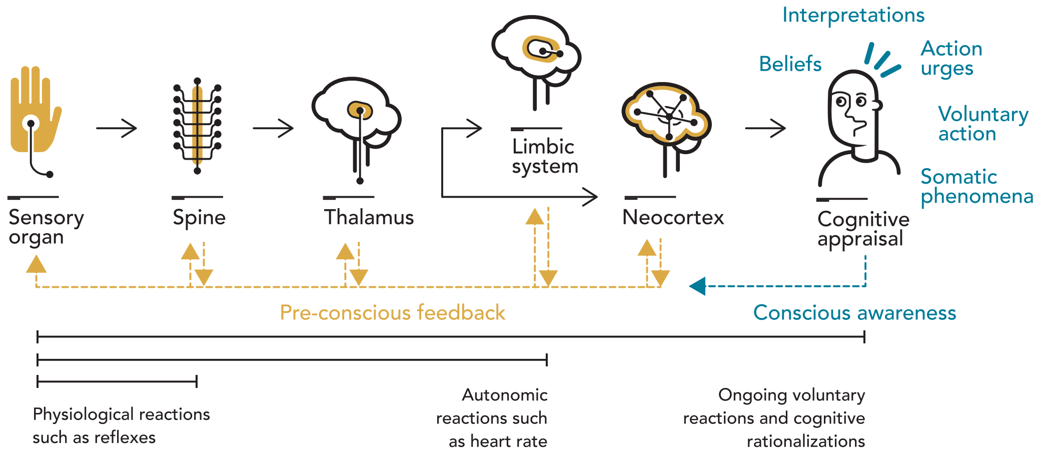


Fig. 1. A spectrum of locations in the body where emotion may be said to occur. Rather than imagining the brain as a singular processing unit with top-down control, it is useful to think of different systems of the body acting at different timescales and with feedback into each other [Parent and Hazrati 1995]. An emotional event will be "experienced" by different parts of the brain differently, each of which is structured and "trained" differently [Sapolsky 2003].

In Figure 1, we illustrate a working map between emotional phenomena and the places in the body where they can be said to "happen." The emotional phenomena that we imagine as singular experiences have a biological basis in different parts of the body and brain. For example, let us examine the emotion of fear. Is fear located anywhere in the brain? In pop culture, people discuss having a "fear center" of the brain. Typically this is rooted in a brain structure called the *amygdala* [Isaacson 2013]. It is called a *fear center* because it is activated when a fast-acting part of the brain called the *thalamus* detects sensory stimuli that have been associated with harm or past fear experiences; then it further activates or inhibits other parts of the brain [Babaev et al. 2018; Davis 1992]. But would it be correct to reduce fear to a single region of the brain? The so-called "fear center" amygdala itself is not actually specific to fear, as it is also involved in processing other emotions, as well as memory [Gainotti 2000]. When the amygdala is disordered or disabled, it does not always result in a deficit in fear [Adolphs et al. 1999]. In fact, there is no one brain region you can disable to cause a specific deficit in a single emotion [Barrett 2012]. Cognition also plays a role in our perception of fear—yet cognition is correlated with a vastly different distributed network of cortical brain regions [Kolb and Taylor 2013]. There is much research on "emotion circuits" or brain networks that give rise to fear [Gainotti 2000; LeDoux 2000; Marek et al. 2013]. But is this sufficient to explain fear? Such an explanation disregards any events related to fear that do not occur in the brain proper, such as increased heart rate and reflexes to sensory events, as well as cultural and sociological context. An adequate explanation of an emotion must examine how it arises in the brain, body, and environment. Let us now change our focus from the brain to the body.

At some level, it is convenient to think of emotions as signals. Nerve signals are responsible for transmitting sensory information to the brain and for controlling muscles and other body parts, and, as far as we know, in some way actually comprise the conscious experience. If we trace a sensation starting from an external event (e.g., a sharp object activating a pain receptor), the signal would pass along nerve fibers to ganglion cells, to the spine, medulla, midbrain, and thalamus, then finally to the amygdala and somatosensory cortex. At each integration juncture, the signal may proliferate other signals (e.g., by instigating a reflex). However, the conscious experience of

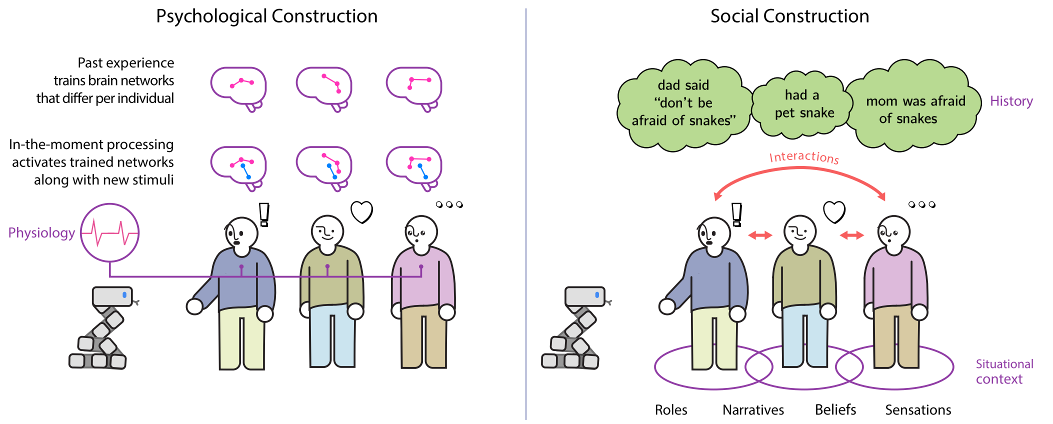

Psychological Construction
Past experience trains brain networks that differ per individual
In-the-moment processing activates trained networks along with new stimuli
Physiology
Social Construction
dad said "don't be afraid of snakes"
had a pet snake
mom was afraid of snakes
History
Interactions
Situational context
Roles
Narratives
Beliefs
Sensations

Fig. 2. Emotion can be psychologically and socially constructed. We integrate past experience, narratives, social relations, and distributed brain networks in understanding our bodily sensations.

the signal would not be possible until the signal has been processed and distributed to the neocortex and amygdala through the thalamus. This means that bodily reactions are already occurring before we are fully conscious of sensation and, further, that multiple parts of the brain will be processing the signals at different times. Much of what we think of as an observable emotional signal are autonomic responses, such as increased GSR, heart rate, or breathing rate. Even if we have some indirect voluntary control over these autonomic responses, the instantaneous reaction is not directly available to our conscious experience; rather, the post hoc sensation of the autonomic response is available. That is to say, we cannot consciously choose to sweat, but we can notice that we have started to sweat after it begins.

Physiological theories of emotion state that our subjective experience of emotion is, at least in some part, either caused by or is exactly the sensation of our bodily reactions to external stimuli. For example, James-Lange theory [Cannon 1987] would state that "feeling angry" is the sum total of feeling your muscles tense and your heart rate increase; two-factor theory would state that the emotional experience is simultaneously partially physiological and partially cognitive. Cognitive appraisal theories state that the cognitive interpretation of an event stimulates the physiological response [Moors et al. 2013]. By contrast, to understand emotions as constructed, it is useful to think of the different parts of the brain and body continually reacting to, being trained on, and processing different data. Psychological construction refers to the interplay between these processes as well as the meaningful portions of the outside world.

3.2 Example of Emotions as Socially and Psychologically Constructed

As an example for the social and psychological construction of emotions, we extend an example from Barrett's work on constructed emotions (Figure 2 presents an accompanying drawing).

Imagine that three people encounter a robot snake. Each will have a different lifetime of experience with robots and snakes, and may have different immediate pre-cognitive reactions. One who was bitten by a snake may be more fearful; another who had a pet snake may be more excited. As their brains process the sensory information, they may have cognitions related to the robot snake that attenuate their immediate reactions. An engineer may recognize the robot as essentially non-lethal and feel calm. A science fiction fan may recognize the robot as something dangerous and feel more fear. These would be examples of immediate individual psychological constructions that have bases in longer-term cultural experiences (social construction). There will be immediate social construction aspects to the encounter as well: they will be continually updating their emotions based on each other's reactions, which may also be inflected by their social status. If the leader of the group is fearful, others may be more worried based on that fear.

If we were to instrument the people with sensors and inspect the data, it is conceivable to observe broadly similar physiological responses from everyone despite their differing emotions: both fear and excitement are correlated with increased heart rate, breathing rate, and GSR. With a granular analysis, we may be able to post hoc reconstruct specific moments of immediate fear, but it is likely that they coincide as much with changes in physical conditions as cognitive rationalizations. The in-the-moment experiences of emotion were affected by past experience and by the shared social experience of observing each other's emotional responses (or lack thereof). Further, all participants in the event later may note that the memory of the experience began to take on more specific meanings as it was recalled and discussed. It would be valid to say that the emotional experience, as filtered through their individual subjectivities, both had in-the-moment differences and post hoc differences as we were able to rationalize and share the narrative of the emotional experience.

3.3 Evidence for Emotions as Socially and Psychologically Constructed

An implication of emotions as socially and psychologically constructed is that each individual's experience of emotional phenomena is highly dependent on their own specific subjectivity, which is itself highly dependent on their interactions with other people both in the moment and over a lifetime. Our subjectivity is (physically) constructed in the brain from a lifetime of experiences where we associate phenomena in the world (e.g., objects, environments, or other people's behaviors) with perceptions of the world and sensations in our body. To use a neuroscience-to-computer science analogy, we can think of construction as being both topological changes in brain networks as well as patterns of activation across brain networks (that also happen to reconfigure the network).

This viewpoint has wide support within psychological emotion research. Ortony and Clore [2013] make the argument for psychological construction based on an evidence-based behavioral and neurological account of the context sensitivity of emotions. Their explanation is that if we understand emotions to be emergent properties of the brain and body as a system, then the context is so highly specific that it is not meaningful to even speak of having consistent experiences for what are labeled as the same emotions. Different structural configurations within the brain between people and the resulting differing cognitive appraisals inflect the experience of emotions.

Similarly, Russell (of dimensional core affect theory [Russell et al. 1989]) has made an argument for the psychological construction of emotions. Although dimensional theories are often operationalized as if affect is something we can easily introspect and determine [Bradley and Lang 1994; Watson et al. 1988], a close reading of Russell's theory posits the affective dimensions of activation and valence to be more like the abstract dimensions of a factor analysis than something we experience consciously [Russell 2003]. A quote from the editorial by Russell [2015] entitled "The Greater Constructionist Project for Emotion" lays bare the level of specificity he believes emotions to have:

> *The concepts of emotion, fear, anger, disgust, and so on are folk concepts that predate psychology. The set of events called emotions, or all those called fear or anger or some other type of emotion, are heterogeneous . . .*

If we take this seriously, labeling emotional experiences with a singular word or point on a scale hides the unique and multifaceted experiential and physiological phenomena occurring during an emotion. This is not to say that we should not try to understand emotions and engage in scientific practices of labeling, categorization and structural modeling but rather that we should approach emotions as socially and psychologically constructed and therefore fundamentally interpretive

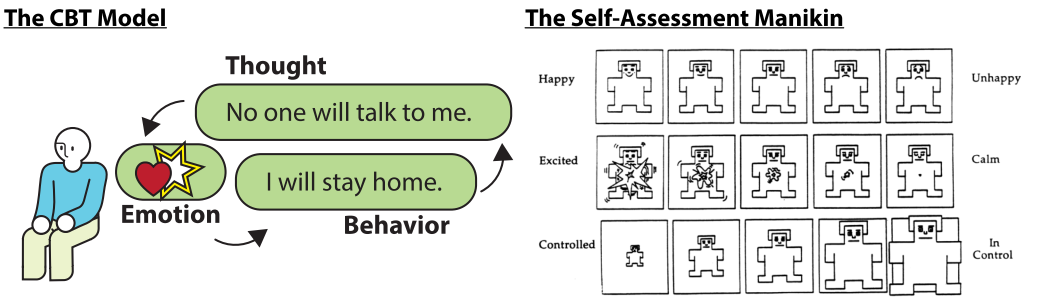


Fig. 3. The CBT model (left) relates thoughts, feelings, and behaviors. The therapeutic concept is that you can intervene on any aspect of the cycle to change your emotions. The point of learning for HRI researchers is that therapists have developed a model such as this because it reflects a common and effective way for people to analyze and communicate about emotions. This is in contrast, for example, to the self-assessment manikin (right), which only inspects an abstracted aspect of "feeling."

phenomena. In other words, our conscious communication of emotion-related phenomena is necessarily dependent on interpretation and representation, based on incomplete introspection.

As an example of the social and psychological construction of emotions, let us first consider the phenomena of *misattribution of arousal*, where a single physiological state (e.g., a heart beating quickly) could be associated with wildly different emotions (e.g., fear, or excitedness) [Cotton 1981]. The classic study by Schachter and Singer [1962] demonstrated this by injecting participants with epinephrine (adrenaline) or a placebo. Of the participants who were given epinephrine, a third were informed of its effects, a third were misinformed, and a third were kept ignorant. They then placed participants in the presence of a happy or angry confederate. They discovered that participants who did not have an adequate explanation of their physiological arousal took on the emotions of their confederate. In this scenario, all participants shared similar physiological states, but their interpretation of those states and resulting constructed emotions differed. Indeed, it is not just social context that can influence emotion but also past experience [Barrett 2017]. In this regard, we can consider social context, cognition, and physiological responses all contributing to emotional experience.

4 THERAPEUTIC APPROACHES AND HOW THEY APPLY TO HRI

Different therapeutic approaches target different aspects of the human experience. In this section, we outline broad therapeutic approaches that we believe HRI researchers can learn from. We do not name every type of therapy even if there may be lessons to be learned for HRI researchers. For example, we do not analyze art and music therapies here. Even though they may teach HRI researchers a lot about emotion expression and the human condition, the way in which to translate their approaches into scientific methodologies is less obvious to us. Further, since therapists are focused on providing effective care, these practices are often mixed and have varied theoretical background. We focus on what HRI researchers can use directly.

4.1 Manualized Therapies

CBT is one of the most widespread therapeutic frameworks [Milne and Reiser 2017]. Emotional interventions are three constituent parts: behaviors, cognitions, and emotions (Figure 3). For example, if a patient believes they are a "bad person," they may engage in behaviors that a "bad person" would do and then might feel guilty, further reinforcing their original belief. CBT aims to intervene on this cycle by asking patients to identify the following: (1) their adverse beliefs (often by writing them down), then rehearsing a countervailing belief; (2) unwanted behaviors and rehearsing alternative behaviors; and (3) unwanted emotions and rehearsing alternative emotions. CBT has been effective at treating a wide variety of mental health difficulties and is heavily *manualized*—that is, CBT relies on manuals, workbooks, and handouts (see the appendix for a DBT manual excerpt) to deliver both psychoeducational content and to help patients practice CBT skills.

DBT draws heavily from CBT [Linehan 2014]; however, it is a holistic intensive training program that is delivered in a simultaneous group and individual format over the course of 6 months to a year. DBT features four modules: distress tolerance, emotion regulation, interpersonal effectiveness, and mindfulness. It is also manualized: group and individual coaches teach 36 skills that patients track their progress in over the course of the therapy. DBT skills are also often taught outside of the core training program through individual therapists, workbooks, and apps. DBT was originally developed to treat borderline personality disorder; however, it has since been used to treat emotional dysregulation corresponding to many diagnoses, including PTSD, depression, and anxiety.

HRI takeaways. Manualized therapies provide ready-made emotion measurement tools that HRI researchers can adopt. They also have extensive accompanying training material.

We mention CBT and DBT as they are common approaches; however. many therapies have been manualized. Importantly, these have been developed and verified through practice and therefore both implicitly and explicitly include methods for grounding the meaning of the materials. For example, the DBT emotion worksheets help a patient label their emotions by providing examples of possible somatic experiences, beliefs, behaviors, and contexts for an emotion. They do not expect a patient to understand the worksheets immediately: the patient works with the group and the coaches over many weeks and months to develop a subtle understanding of each emotion (see the appendix).

Perhaps unsurprisingly, CBT and DBT take a mostly cognitivist approach to emotions—that is, expecting that humans experience categorical emotions, training people in differentiating emotions, and emphasizing the importance of intervening on cognitive beliefs. DBT more explicitly grounds emotion in the body through training in mindfulness that includes bodily awareness.

4.2 Somatic Therapies

Somatic therapies are mostly focused on the bodily (somatic) feeling and expression of emotions [Barratt 2010; Van der Kolk 1994, 2015]. They aim to develop a patient's conscious awareness of the somatic experience of an emotion and to develop body-based emotional interventions. Somatic therapies are guided by an ontological principle of embodied emotion; in contrast to other therapies that focus on narrative and/or cognition, a somatic approach focuses on the physical extent of emotional trauma as it is encoded in the body/nervous system(s). Emotional experience is expressed via inarticulable modalities such as physical movement and touch. These are augmented by associating localized body sensations with sensory metaphors (e.g., "hot," "red," or "sharp" for sense of emotional pain).

HRI takeaways. The key insight for HRI researchers is how somatic therapies focus on body movement, localization, and metaphor to describe emotion experiences.

Rather than assuming that an emotion is easy to identify and label, the fundamental assumption of somatic therapy is that many different metaphors and associations are needed to explicate an emotional experience. Further, there is a strong conceptual link to HRI: it is common for HRI researchers to be interested in gestures, touch, and personal space, or to instrument the participant's body with sensors. The somatic assumption that emotion is encoded in, produced by, and expressed through the body is entirely compatible with physically grounded HRI studies. HRI could benefit from techniques to gain shared understanding of emotions (epistemology) of somatic therapy.

4.3 Narrative Therapies

When people think of therapy, they often think of talk therapy as Freudian psychoanalysis. Although the field has developed in the almost-century since Freud, talk therapies are still the basis of many other approaches; practitioners will often incorporate many other approaches (e.g., somatic, DBT) into their talk therapy sessions.

Narrative approaches focus on the cognitive and memory aspect of emotions [Madigan 2011]. Ontologically, they are quite cognizant of a person's emotions and behaviors being the product of years of experience, and often seek to locate the narrative origin of current emotional difficulties. Epistemically, they use the method of storytelling to develop a patient's self-understanding. Some are quite explicit in their storytelling methods. For example, family constellation therapy asks a group of people to literally act as characters of a target patient's family so that they may theatrically perform healing moments. Sandbox therapy asks a patient to associate memories, emotions, and self-conceptions with arrangements of toys in a sandbox (and is often used to help children express trauma). Therapists are quite involved in the narrative development and act as a guide or interpreter for the patient's narrative experience.

Narrative therapies that take a post-modern approach seek to analyze a patient's experience in terms of cultural narratives. For example, feminist narrative therapy will work to develop a patient's understanding of their identity in relation to cultural scripts and meta-narratives, then try to "rewrite the script" for the patient.

HRI takeaways. Narrative can determine how a participant receives, conceptualizes, and reports on an emotional experience. Narrative therapies provide techniques for managing and grounding these narratives and could be used by HRI researchers.

Particularly for studies in which a robot is presented as a social actor, the narrative that the participant develops about the robot can entirely determine their emotional perception of the robot. This is evident in several HRI studies [Bucci et al. 2018; Jung and Hinds 2018; Ling and Bjorling 2020; Marino et al. 2017] that have studied narrative's impact on emotional ratings of robots. Even if HRI researchers would prefer to ignore the interpretive elements of narrative interaction, it is obvious that participants will engage in narrative interpretation whether or not the researchers would like them to.

4.4 Trauma-Informed Approaches

"Trauma-informed care" is used by different communities to mean different things. As a result, it can be confusing to understand what it refers to. For the purposes of this article, we take trauma-informed approaches to care to mean an ethical stance that prioritizes the agency of the patient above all else, and the resulting ethic of care that prioritizes careful consideration of what might be emotionally triggering for patients to experience [Raja et al. 2015].

A trauma-informed approach can be used in any care-providing service, from healthcare to psychotherapy to immigration support services and more. The guiding principle for trauma-informed care is that the person who receives the care (patient/client) should be in total control of the care that they receive. The insight is that people who have suffered traumatic experiences, whether physical or emotional, have lost a sense of agency over their lives that needs to be preserved/redeveloped. As such, trauma-informed care is more of a statement about a power relation between the care providers and the patient/client: the institutional positionality fundamentally puts them in a position of power over their client, and they need to actively work to subvert that power relation by handing control over to the client.

For example, in an emergency ward, a doctor is institutionally empowered to decide the kind of treatment that a patient will receive. Even the most ethical doctor cannot change this institutional power: it is not a moral statement but just a fact of the structure of the hospital that the doctor controls the patient's care. This is because the patient (1) does not know about all of the types of care that are possible, (2) does not have the same institutional access to their own data as their doctor (e.g., a patient must file a request to get their own medical records), and (3) is unable to requisition their own medical procedures (the patient cannot get an x-ray without the doctor making the request).

A critical look at the institutional relations of the hospital would point out that people who have particular identities are often denied the kind of care that they need as a result of these kinds of power imbalances. For example, endometriosis is often not correctly diagnosed as a result of doctors who do not take women's expressions of menstruation pain to be serious enough to warrant medical examination; simultaneously, women are typically trained to express pain in different ways due to cultural narratives about menstruation [Samulowitz et al. 2018]. A trauma-informed approach would instead allow a woman who is experiencing pain decide for herself how serious it is and instead facilitate the kinds of care that she thinks is necessary by providing medical knowledge and discussing options. Further, the hospital would try to provide ways to intervene on yet-unknown harms by establishing care procedures that account for possible trauma, creating patient-led advisory boards to change hospital practices, and strengthen accountability and grievance resolution processes [Raja et al. 2015; TICIRC 2020].

HRI takeaways. The lesson for HRI researchers from trauma-informed care is to acknowledge the structural power that they have over participants.

This is not to say that HRI researchers are necessarily in the same position as doctors in terms of being able to deny care. Researchers have structural power because they provide the study materials and environments that fully determine a participant's experience in the moment during a study and afterward in terms of analysis and reporting. We suggest that HRI researchers attempt to prioritize the participant's agency during a study, including designing ways for a participant to be emotionally safe while a study procedure is occurring, and for the participant to be able to give feedback about study procedures.

5 ACCOUNTING FOR SUBJECTIVITY IN HRI STUDY DESIGNS

In this section, we (1) conceptually address the four errors outlined in Section 2.1 and (2) provide worked examples of how we imagine HRI researchers could work with therapists to solve those errors. We provide illustrated examples to explain our position in the text and have included worksheets in the appendix.

5.1 Addressing Categorical Errors

The source of categorical errors lies in a mismatch between the experiential realities of emotion and the measurement, perpetuated by emotion theories that obfuscate the constructed nature of emotions. Robots are interactive, so our argument is simple: we should learn from the people who interact with emotions daily to develop a theoretical approach that is appropriate for interactive computational agents. The ontological statement that emotions are socially and psychologically constructed means that emotional phenomena are much more complex than we often account for in our study designs. The epistemic claim is that therapists have the practical expertise in how to draw out/capture other people's emotions. Further, claims with regard to a robot's therapeutic

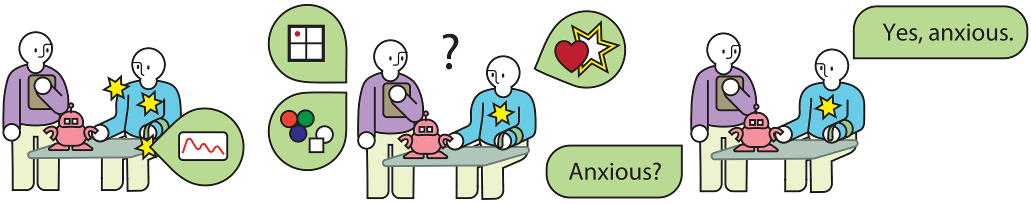


Fig. 4. In this figure, we see the researcher (purple) and participant (blue) engaged in a GSR trial related to stress. The yellow stars indicate the participant's many somatic experiences, only one of which may correspond to the GSR graph. Further, without discussion, we do not know what the participant is experiencing as their stress response, or which aspects of the stress response are being captured in the GSR. Clarifying this would help us understand the categorical differences between stress as measured vs. stress as experienced. Then, we see the researcher asking the participant to represent their experience with a common 2D affect grid and using shape and color metaphors to expand their shared understanding. Last, we see the participant and researcher agree to call that somatic experience "anxious," which can serve as a grounded term for the rest of the study.

benefit should be reviewed by a real therapist who will have the experiential knowledge to "gut-check" claims.

Concretely, we recommend that HRI researchers avail themselves not only of emotion theory but also common practices of therapists. *We do not recommend that HRI researchers become therapists.* Rather we suggest that there is much practical knowledge that could be leveraged to clarify emotional concepts and measurement in a study design. This is similar to the common practice of hiring statisticians to assist with study design and analysis: we believe that emotion research requires specialized knowledge to apprehend (and respect) the complexity of human emotional experiences. Consulting with therapists (or hiring them to do data collection) can provide the critical perspective necessary to understand which *category* of emotion phenomena we are attempting to study and whether our methods are appropriate.

Takeaway 1: *We do not recommend that HRI researchers become therapists but instead recommend hiring therapists to review methods and assumptions about emotional phenomena. Like statisticians, they are practical experts in their field—namely, emotion elicitation and analysis. If a study claims to have a therapeutic benefit, the results should be verified by a therapist.*

Example: Autonomic responses. Autonomic responses can be measured with electronic sensors and are often used to determine the participant's emotional state. For example, researchers may want to determine the participant's stress level through GSR and automatically apply stress reduction interventions. Often, GSR is used as a direct proxy for stress.

However, if we view stress as constructed, we would have to account for the ongoing simultaneous but categorically different factors that comprise the stress experience. We would want to account for and differentiate the participant's immediate evolving somatic experience from their cognitive rationalizations, which means accepting that a participant can only communicate with limited available language. If we had to capture a "stress level," we would commit to spending time with the participant to substantiate which of the categorically different parts of the stress experience we would like them to introspect about (see Figure 4).

Hiring a therapist to consult on study design would give the researcher options and clarify the categorically different stress phenomena that would be apprehended via GSR vs. via a somatic approach. Simply put, we suggest a professional gut-check: a therapist has practical expertise to know what category of emotion is being inspected.

Table 1. List of Phenomena of Interest That Comprise an Emotional Event Inspired by a Framework from DBT [Dimeff and Linehan 2001]

O/I	Phenomena	Explanation	Examples
O	Autonomic reactions	Bodily responses that occur pre-attentively	Sweating, heart rate
O/I	Expressions	In-the-moment actions that can be controlled attentively	Facial expressions, gestures, intentionally slowing breath
O/I	Behaviors	Longer-term actions actually undertaken by a person	Actually crying or running away from robot
O/I	Prompting events	Rationalized causes for emotional reactions	Deciding that falling down made me sad
I	Somatic experience	Sensations felt in the body or brain	Muscle tension, headache, warmth
I	Narrative framing	Rationalization of interaction in terms of roles and scenarios	Deciding that the robot is a “nurse”
I	Action urges	Actions a person may want to do	Wanting to cry/run away from the robot
I	Interpretations	Guesses at others’ feelings or consequences	Deciding that the robot is “happy”
I	Beliefs	Generalized statements about self or others, could be metaphorical “suspension of disbelief” or “true beliefs”	Deciding the robot cannot actually feel emotion
I	After effects	Interpretations, actions, beliefs, and somatic experiences that occur after an event is “over”	Noticing a lingering tension for some time after a robot has scared you

Note: Objective phenomena (O) have causally observable quantities that can be measured and compared using physical sensing equipment (sensors, rulers, etc.). Interpretive phenomena (I) require some adjudication through language and introspection. Behaviors and expressions are differentiated here by duration and level of attention—that is, a behavior requires at least some attentive voluntary control, but expressions may or may not require attentive control. The objects and actions within events are observable, but categorization requires some interpretation. See the appendix for more examples.

This is not an entirely new suggestion. HRI researchers often work with domain experts to differentiate scientific/engineering claims from claims that require rigor within the humanities. For example, Park et al. [2019] employed experts in literacy to assist with their literacy robot and were deployed in schools; Wood et al. [2019] employed teachers who worked with children who have autism to ground their work.

Somatic examination may reveal that stress was phenomenologically different enough between participants to be a meaningfully different kind of emotional response, which researchers would want to account for in post hoc analysis.

Drawbacks. Besides the obvious difficulty in adopting new theory for researchers, the main difficulty for this approach would be the implications for study design and implementation cost. Theory drawn from somatic therapy is not well substantiated in HRI literature and common HRI-related psychological sources. Common validated methods would have to be reconsidered.

5.2 Addressing Methodological Errors

In contrast to how we expect our study participants to be able to make on-the-spot emotional assessments, therapists usually train clients over long periods of time to introspect and determine a variety of emotional phenomena. In an interaction, there is a subjective interplay between beliefs, behaviors, and bodily sensations. Each therapy assumes that introspection requires a therapist to train and practice with their clients to determine a variety of emotional phenomena (Table 1). This is in contrast to implicit assumptions in HRI studies that participants should be able to “dead-reckon” their emotions without much training. We argue that HRI researchers should form their methodologies with the principle that *emotions are difficult to introspect accurately.*

Methodological errors occur when this principle is violated. However, it is understandable that it would be violated because of practices within academic psychology that reasonably try to limit the impact of researcher bias. For example, it is common to treat participants as “blank slates” and provide “validated” surveys and treatments as if they are neutral experimental factors. For example, the Positive And Negative Affect Schedule (PANAS) is a “validated” mapping of emotion words to affect grid quadrants—or libraries that map movie clips to emotion ratings [Gabert-Quillen et al.

2015], the assumption being that it is a standard treatment factor that can be applied to produce a particular emotion in a participant.

For both of these "validated" scales, the implicit assumption is this: if there are deviations in participant understandings of the mapping between words and affect grid quadrants in *your study*, then they will be normally distributed and accounted for by the central limit theorem during analysis.

A constructed view of emotion would entail that we expect each participant's experience of an emotional phenomenon to be different. Further, we would imagine that their reaction would be sensitive to conditions. As such, we would not know whether these validated emotion scales and factors are, in fact, producing or capturing the subjective experience we expect. We would have to rely on the quality of participant introspection to trust our measurements.

As such, we recommend that (1) participants are trained in introspective methods and (2) measurements are triangulated by approaching each emotion as a combination of somatic, cognitive, and narrative aspects. A participant should be made aware of the meaning of an emotion measurement by training them in each differentiated emotion. This may be easier than it sounds: manualized therapies provide robust frameworks for this.

Takeaway 2: *We should train our participants in noticing what is happening in their bodies. Emotions are hard to measure by cognitive introspection, which takes years of practice to develop.*

Example: Emotion training for "guilt" vs. "shame." The DBT manual has emotion sheets that can be directly used by HRI researchers that explain the full experience of an emotion. See the appendix for examples: guilt and shame are chosen as illustrative examples because these are often difficult for a person to differentiate. An HRI researcher would go through the manual step by step with the participant to ground their experience.

The manual specifies prompting events—that is, which events would reasonably make someone feel "guilt" or "shame." To help differentiate, a researcher would read through the events with the participant and then ask, "Can you think of events that are like this that made you feel guilty?" Depending on the participant's response, the researcher would either confirm or amend the participant's response (say: "Ah, we think of that more as shame than guilt."). Each item of the manual would be explained in a similar way: common body experiences, beliefs, behaviors, and related emotion words. Then each emotion word would be substantiated in terms of the participant's own experience—grounded—and differentiated according to the researcher's intended study factors.

The introspective training would provide the researcher with important insights that they would use to ground the rest of their measurement and discussion. Grounding in agreed-upon insights resolves ambiguities that may be present in the participant's experience of the emotion.

Drawbacks. The preceding process would add time to the study and require training for the researcher. However, the stronger critique is that it introduces researcher bias into the study. This could become problematic in larger-*N* studies with many different research assistants who run participants, presenting a greater need for consistency controls. The process also excludes non-in-laboratory surveys as a possible method since it requires iterative feedback.

5.3 Addressing Instrumental Errors

A major assumption of HRI emotion research is that emotions can be labeled with words and scales that meaningfully describe the subjective experience of an emotion. However, the view that emotions are constructed would imply that we should make these words and scales meaningful to each individual who attempts to reason with them. Effectively, we would have to co-construct a scale with a participant by training them in our scale's meaning through reference to their own experience (similar to the preceding). This would include (1) familiarizing the participant with our definitions of the somatic experience of particular emotions, (2) asking for the participant to benchmark certain words by describing their memories of a particular emotion, and (3) helping the participant to identify in themselves the difference between scale items (e.g., what is the difference between a 2/5 level of guilt and a 3/5 level of guilt?).

We argue that turning to somatic therapy for guidance would help here. Somatic therapists specialize in using multiple metaphors to address the in-the-moment experience of an emotion. A somatic therapist might ask about metaphors such as the shape, color, or hardness of a sensation, working with a client to develop the client's understanding of their own sensations.

This has precedence in pain management [Rosier et al. 2002]. Pain is understood as a highly personal experience: someone's previous experiences of pain affect their current experience of pain, and cognitive beliefs relating to their pain are known to impact the emotional processing of that pain [Lamé et al. 2005]. As such, doctors will administer pain measurement scales in a way that benchmarks the scale by asking the person to imagine the most and least pain that they have experienced to ground the meaning of a "10"and a "0" [Ong and Seymour 2004]. Studies in symptomatology incorporate metaphors to help a patient describe the experience of their pain (e.g., a sharp pain or a throbbing pain), which can aid in diagnosis or therapeutic reconceptualization [Gallagher et al. 2013]. Studies that attempt to aggregate pain measurements across patients have to account for this individual variability [Manworren and Stinson 2016]. Further, it is understood that the act of measuring can often heavily influence the outcome of the scale measurement. For example, one study showed a large discrepancy between the amount of pain patients reported on paper scales administered by nurses in person vs. electronic scales administered remotely [Price et al. 2018]. The important lesson with scales for pain management is that even if we assume some universal mechanism for sensing pain, the perceptual aspect may be significantly different due to different past experiences that our brain was exposed to. Further, how we express and describe pain is influenced by our beliefs, ability to remember past pain, the social dynamics of the measurement process, and our understanding of the meaning of the scale.

This does not mean that we cannot use scales, but that we should understand that scales that measure subjectivity are necessarily relative to a person's experience. Despite the variability between patients in therapy programs, therapists often still make heavy use of scales. Similar to pain management, these scales are understood to be relative to the patient's own experience.

Takeaway 3: *Scales are relative to a person's experience, but that does not make them scientifically useless. Instead, we need to benchmark them to the participant's own experiences.*

Example. We imagine that a researcher would discuss with a participant the methods for attending to sensations in their bodies and work to co-develop metaphorical representations of the sensation. Say that in this case we were inspecting "fear." The researcher may ask the participant to recount a fearful event. Then, they would ask "Where in the body does the fear express itself for the participant?" The participant may answer "as tears," or "in my chest." The researcher would then ask the participant to substantiate the sensation with a metaphor, offering examples of colors ("Is the fear blue or yellow?"), shapes ("Is the fear sharp or round?"), textures ("Is the fear rough or soft?"), temperatures ("Is the fear hot or cold?"), and so on. Then the researcher would ask the participant to benchmark their fear responses to scale items, such as a 2/3 fear is "hot,"but a 3/3 fear is "cold."This provides metaphors that are more commonly used on scales and therefore can be reasoned about between participants.

Drawbacks. For within-participant designs, using different metaphors to substantiate the scale may make the scale inconsistent with certain statistical techniques.

For example, it is an ongoing debate within quantitative psychology as to the validity of treating Likert scales as continuous linear variables [Pimentel and Pimentel 2019]. We can understand why HRI researchers would prefer to treat them that way, particularly for regressions. It also complicates between-participant analysis: if one person's baseline is different than another's, or one person's metaphor is different from another's, can we validly group them during post hoc tests? Incorporating metaphors into analysis could therefore decrease statistical power by virtue of having more blocks, factors, or groups.

5.4 Addressing Complexity Errors

We contend that due to the constructed nature of emotions, it is best to think of that which is expressed during a study or captured during a measurement as highly unstable. In a complex system such as an emoting human, there are many hidden variables and/or processes that can impact any specific expression. In the preceding, we have addressed resolutions to measurement ambiguity, starting from conceptual/categorical clarity to methodology and study instruments. Here we address the ontological vs. epistemic concerns of what emotions are vs. how they are expressed.

Since robots are often situated as social agents, studies need to account for emotions being the product of a process of in-the-moment experiences. Social science provides us with frameworks for understanding certain social dynamics that may be at play within our studies that are difficult to expose. This section offers theoretical frameworks to guide behavior analysis with reference to social systems.

Taking a feminist narrative therapeutic perspective would suggest looking at emotional interactions from a critical narrative lens by situating the participant and robot relative to the participant's self-understanding and perceived power dynamics. Dramaturgical theories of emotion align with certain critical feminist perspectives, as emotional expressions are assumed to be fundamentally performative. A study from this perspective would examine the conflict between a robot's intended displayed emotion, actually communicated emotion, and internally felt emotions. Behavior from this perspective is thought to be representational of internal states, but abstracted and mediated through identity and social norms.[2] In the view of dramaturgical emotion theory, what is expressed during the interaction has a different emotional tenor than the subjective experience of each person alone. In affective robotics terms, assigning an emotional label to the behaviors would not give the only reading of someone's internal emotional experience, but instead what they felt it was appropriate to convey [Turner and Stets 2006]. Symbolic interactionist theories center the reinforcement of one's own self as the primary objective for emotional motivation, where identity may include multiple, overlapping identities [Stets and Turner 2014]. Symbolic interactionists imagine emotion as a continuous process that produces and also results from identity. Identity is continually negotiated with regard to cultural norms, beliefs, and social roles.

Feminist narrative therapy addresses an individual's relationship to cultural "meta-narratives" by incorporating social science theory directly into the therapeutic process. For example, someone may examine their own relationship to common cultural understandings of gender and attempt to "rewrite" their personal belief systems relative to these cultural narratives. For example, if someone who identifies as a man feels that they are "not strong enough to be a man," a feminist narrative therapeutic approach would encourage them to rewrite their own narrative of what it means to "be a man" rather than try to "become stronger."

An HRI approach that uses therapeutic practices founded in social science theory can help address complexity errors because of the awareness that social theory can bring to often-unseen cultural forces. They can help expose hidden variables, provide language for roles/responsibilities/beliefs that are impacting a participant's emotions, or serve as a theoretical basis for analysis. In

[2] For example, one might perform being more upset at something someone says to them during a meeting than they may truly feel for the goal of adhering to group norms or garnering group sympathy.

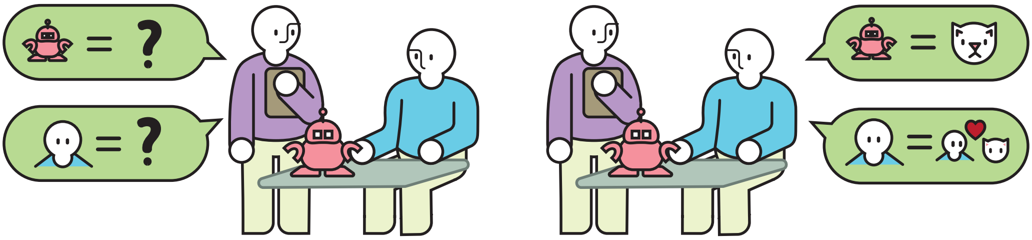


Fig. 5. In this figure, we see the researcher speaking with the participant about how they see the robot within a narrative. By grounding the robot with a narrative such as "the robot is like my cat," both the participant and the researcher have established the boundaries of the "suspension of disbelief" that is necessary to see a robot as an autonomous agent. It further explicates the participant's emotion reactions within an accessible cultural frame. The emotional relationship between cats and owners is known as cultural concepts that provide common referents.

the previous example, the therapist is able to use a feminist framework to go beyond a surface-level understanding of their patient's emotional experience by exploring the sociological factors that shape it.

Takeaway 4: *There are many social theories that can provide us ready-made frameworks for addressing social complexities. Not addressing them does not make the impact of social concepts go away, it just means that we have not accounted for them in our study designs.*

Example. During a study, an HRI researcher would try to address which kinds of social dynamics might be at play. The robot's "story" can be provided by the researcher or built with the participant. The social role of a robot can drastically change participant perceptions of emotional behaviors [Chen et al. 2020]. Whether or not we intend it, robots can be seen by participants as existing in a make-believe world (Figure 5). Narrative therapies and interpretive emotion theories would provide insight into how to help resolve this.

A *feminist narrative perspective* would encourage critical reflection as to how robots are integrated into systems of power. For example, a teaching robot would be examined in terms of the role of a teacher in producing emotions as opposed to the effectiveness of administering information (see Figure 5 for an illustration of roles).

A *dramaturgical approach* views robot behavior as performance, which would engender questions about the robot's role in an interaction and would encourage critical reflection of the congruity between the internal states and externally expressed states of interactors. For example, a participant may believe that a robot is masking a true "hidden' emotion with a smile.

Finally, viewing robots through a symbolic interactionist lens would call into question how the interactors are reinforcing cultural norms through their behaviors, and how that affects the identity of all parties. For example, a participant might believe that a dog-shaped parking enforcement robot is acting in the role of the police due to the employment of dogs in the police force.

Drawbacks. Qualitative and interpretive methods are harder to analyze and easy to misuse. HRI researchers are used to mixed methods, but there is always a question of establishing rigor and reproducibility. This is difficult to adjudicate or convey through writing, as interpretive methods require high levels of skill to administer well and offer few objective measures of success (e.g., it is hard to know whether an interview was done "well" or whether a study's success rests on a researcher's ability to create rapport with a participant). Along with that comes training in the methods and analytical approaches of each theory. For example, rigorous qualitative analysis usually requires stating philosophical positionality so that readers know which philosophical framework is being applied. It can also be difficult to mix theoretical approaches due to apparent philosophical incompatibilities.

6 DISCUSSION

In this article, we have presented the position of emotion theorists who view emotions as psychologically and socially constructed. In taking this position, we have made the argument that HRI researchers can learn from therapeutic practitioners to capture more of the full picture of constructed emotions. In particular, we have presented four errors that we believe can be resolved by learning from therapeutic approaches. These errors focused largely on the act of in-the-laboratory measurement, from theory that would impact study design to the actual carrying out of study procedures. However, there are many other kinds of errors that we did not mention. For example, we did not talk about internal or external validity, which could be threatened by untested new methods. Similarly, ecological validity is a particularly pressing concern for HRI researchers who want to create laboratory environments that are microcosms of prospective real-world environments. Particularly as robots proliferate in human environments, questions about the real ongoing embodied experience with robots become more pertinent.

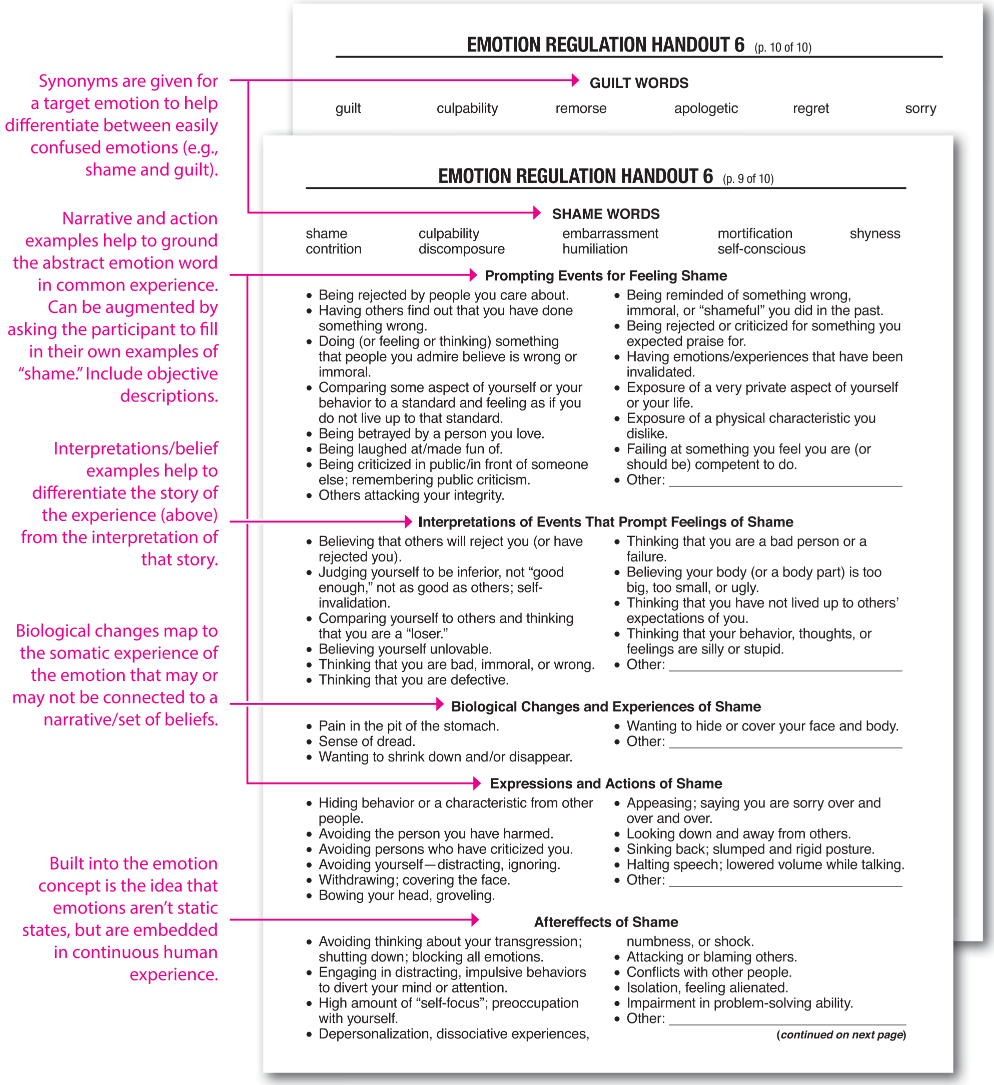
We are cautious about presenting our work as if it is particularly invalidating previous work. We prefer to think of it as growing the nuances and complexity of the subtle art of emotional interaction along with the science. For us, there is explanatory power in our approach, shedding light on the questions of why is it so difficult to reliably create emotional experiences with robots, and why it is so difficult to contain those emotional experiences in a scientific inquiry. It is our hope that this work is used to explicate other researchers' own feelings of dissatisfaction with study methods that engender questions of emotional validity. That is where this work came from for us—that is, in fundamentally asking and answering for ourselves: how can we know whether our studies are getting at the phenomena we purport to be inspecting?

Last, we believe that some of these changes are more of a matter of starting from a different perspective rather than a complete methodological overhaul. We use the methods that we do for good reasons: mostly in a rigorous attempt to manage bias and make sense of complex phenomena. Adopting the constructed view of emotions presents a starting point for understanding emotions as embedded in complex systems; using therapeutic methods may allow us to import the practical knowledge of those who do emotion understanding in their daily work.

7 CONCLUSION

We have presented a working understanding of the socially and psychologically constructed nature of emotions and the implications for affective robotics theory and methodology. We concluded that knowing definitively that robot bodies and behaviors will evoke certain emotions is methodologically questionable. We propose that, beyond simply looking to qualitative constructivist methods, we can learn from therapeutic practices. Therapeutic practices are especially relevant for embodied affective robotics because they have been developed over years by practitioners experienced in developing subjective emotional understandings with clients. We believe that adopting these ways of understanding emotion can produce a paradigmatic shift in affective computing methodologies wherein specific emotional phenomena can be targeted.

APPENDIX



EMOTION REGULATION HANDOUT 6 (p. 10 of 10)

GUILT WORDS

guilt culpability remorse apologetic regret sorry

EMOTION REGULATION HANDOUT 6 (p. 9 of 10)

SHAME WORDS

shame
contrition
culpability
discompose
embarrassment
humiliation
mortification
self-conscious
shyness

Prompting Events for Feeling Shame

- Being rejected by people you care about.
- Having others find out that you have done something wrong.
- Doing (or feeling or thinking) something that people you admire believe is wrong or immoral.
- Comparing some aspect of yourself or your behavior to a standard and feeling as if you do not live up to that standard.
- Being betrayed by a person you love.
- Being laughed at/made fun of.
- Being criticized in public/in front of someone else; remembering public criticism.
- Others attacking your integrity.
- Being reminded of something wrong, immoral, or "shameful" you did in the past.
- Being rejected or criticized for something you expected praise for.
- Having emotions/experiences that have been invalidated.
- Exposure of a very private aspect of yourself or your life.
- Exposure of a physical characteristic you dislike.
- Failing at something you feel you are (or should be) competent to do.
- Other: ____________________

Interpretations of Events That Prompt Feelings of Shame

- Believing that others will reject you (or have rejected you).
- Judging yourself to be inferior, not "good enough," not as good as others; self-invalidation.
- Comparing yourself to others and thinking that you are a "loser."
- Believing yourself unlovable.
- Thinking that you are bad, immoral, or wrong.
- Thinking that you are defective.
- Thinking that you are a bad person or a failure.
- Believing your body (or a body part) is too big, too small, or ugly.
- Thinking that you have not lived up to others' expectations of you.
- Thinking that your behavior, thoughts, or feelings are silly or stupid.
- Other: ____________________

Biological Changes and Experiences of Shame

- Pain in the pit of the stomach.
- Sense of dread.
- Wanting to shrink down and/or disappear.
- Wanting to hide or cover your face and body.
- Other: ____________________

Expressions and Actions of Shame

- Hiding behavior or a characteristic from other people.
- Avoiding the person you have harmed.
- Avoiding persons who have criticized you.
- Avoiding yourself—distracting, ignoring.
- Withdrawing; covering the face.
- Bowing your head, groveling.
- Appeasing; saying you are sorry over and over and over.
- Looking down and away from others.
- Sinking back; slumped and rigid posture.
- Halting speech; lowered volume while talking.
- Other: ____________________

Aftereffects of Shame

- Avoiding thinking about your transgression; shutting down; blocking all emotions.
- Engaging in distracting, impulsive behaviors to divert your mind or attention.
- High amount of "self-focus"; preoccupation with yourself.
- Depersonalization, dissociative experiences, numbness, or shock.
- Attacking or blaming others.
- Conflicts with other people.
- Isolation, feeling alienated.
- Impairment in problem-solving ability.
- Other: ____________________

(continued on next page)

Synonyms are given for a target emotion to help differentiate between easily confused emotions (e.g., shame and guilt).

Narrative and action examples help to ground the abstract emotion word in common experience. Can be augmented by asking the participant to fill in their own examples of "shame." Include objective descriptions.

Interpretations/belief examples help to differentiate the story of the experience (above) from the interpretation of that story.

Biological changes map to the somatic experience of the emotion that may or may not be connected to a narrative/set of beliefs.

Built into the emotion concept is the idea that emotions aren't static states, but are embedded in continuous human experience.

Fig. 6. An excerpt from the DBT manual [Linehan 2014] for emotion words. This can be directly used by HRI researchers if shame is meant to be studied or adapted for emotions of interest. Provided by contrast for guilt (above), as these are emotions that are commonly confused and may be useful to differentiate. Emotion sheets such as these can give context to emotions and can help ensure that participants have grounded concepts by which they can differentiate their self-measurements.

REFERENCES

Ralph Adolphs, Daniel Tranel, S. Hamann, Andrew W. Young, Andrew J. Calder, Elizabeth A. Phelps, Al Anderson, et al. 1999. Recognition of facial emotion in nine individuals with bilateral amygdala damage. *Neuropsychologia* 37, 10 (1999), 1111–1117.

Olga Babaev, Carolina Piletti Chatain, and Dilja Krueger-Burg. 2018. Inhibition in the amygdala anxiety circuitry. *Experimental & Molecular Medicine* 50, 4 (2018), 1–16.

Barnaby Barratt. 2010. *The Emergence of Somatic Psychology and Bodymind Therapy*. Springer.

Lisa Feldman Barrett. 2012. Emotions are real. *Emotion* 12, 3 (2012), 413.

Lisa Feldman Barrett. 2017. The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience* 12, 1 (2017), 1–23.

Lisa Feldman Barrett and James A. Russell. 2014. *The Psychological Construction of Emotion*. Guilford Publications.

Howard Bath. 2008. The three pillars of trauma-informed care. *Reclaiming Children and Youth* 17, 3 (2008), 17–21.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59.

Paul Bucci, Lotus Zhang, Xi Laura Cang, and Karon E. MacLean. 2018. Is it happy?: Behavioural and narrative frame complexity impact perceptions of a simple furry robot's emotions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 509.

Paul H. Bucci, X. Laura Cang, Hailey Mah, Laura Rodgers, and Karon E. MacLean. 2019. Real emotions don't stand still: Toward ecologically viable representation of affective interaction. In *Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII'19)*. 1–7. https://doi.org/10.1109/ACII.2019.8925534

David Byrne and Gillian Callaghan. 2013. *Complexity Theory and the Social Sciences: The State of the Art*. Routledge.

Walter B. Cannon. 1987. The James-Lange theory of emotions: A critical examination and an alternative theory. *American Journal of Psychology* 100, 3–4 (1987), 567–586.

Huili Chen, Hae Won Park, and Cynthia Breazeal. 2020. Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education* 150 (2020), 103836.

Gerald L. Clore and Andrew Ortony. 2013. Psychological construction in the OCC model of emotion. *Emotion Review* 5, 4 (2013), 335–343.

John L. Cotton. 1981. A review of research on Schachter's theory of emotion and the misattribution of arousal. *European Journal of Social Psychology* 11, 4 (1981), 365–397.

Michael Davis. 1992. The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience* 15, 1 (1992), 353–375.

Linda Dimeff and Marsha M. Linehan. 2001. Dialectical behavior therapy in a nutshell. *California Psychologist* 34, 3 (2001), 10–13.

Crystal A. Gabert-Quillen, Ellen E. Bartolini, Benjamin T. Abravanel, and Charles A. Sanislow. 2015. Ratings for emotion film clips. *Behavior Research Methods* 47, 3 (2015), 773–787.

Guido Gainotti. 2000. Neuropsychological theories of emotion. In *Neuropsychology of Emotion*, J. C. Borod (Ed.). Oxford University Press, 214–236.

Laura Gallagher, James McAuley, and G. Lorimer Moseley. 2013. A randomized-controlled trial of using a book of metaphors to reconceptualize pain and decrease catastrophizing in people with chronic pain. *Clinical Journal of Pain* 29, 1 (2013), 20–25.

Per Holth. 2001. The persistence of category mistakes in psychology. *Behavior and Philosophy* 29 (2001), 203–219.

Robert Isaacson. 2013. *The Limbic System*. Springer Science & Business Media.

Malte Jung and Pamela Hinds. 2018. Robots in the wild: A time for more robust theories of human-robot interaction. *ACM Transactions on Human-Robot Interaction* 7, 1 (2018), Article 2, 5 pages.

Malte F. Jung. 2017. Affective grounding in human-robot interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. 263–273. https://doi.org/10.1145/2909824.3020224

Bryan Kolb and Laughlin Taylor. 2013. Neocortical substrates of emotional behavior. In *Psychological and Biological Approaches to Emotion*. Psychology Press, 133–162.

Inge E. Lamé, Madelon L. Peters, Johan W. S. Vlaeyen, Maarten V. Kleef, and Jacob Patijn. 2005. Quality of life in chronic pain is more associated with beliefs about pain, than with pain intensity. *European Journal of Pain* 9, 1 (2005), 15–24.

Lucian Leahu and Phoebe Sengers. 2014. Freaky: Performing hybrid human-machine emotion. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS'14)*. 607–616. https://doi.org/10.1145/2598510.2600879

Joseph E. LeDoux. 2000. Emotion circuits in the brain. *Annual Review of Neuroscience* 23, 1 (2000), 155–184.

Marsha Linehan. 2014. *DBT Skills Training Manual*. Guilford Publications.

Honson Y. Ling and Elin A. Bjorling. 2020. Sharing stress with a robot: What would a robot say? *Human-Machine Communication* 1 (2020), 133–159.

Stephen Madigan. 2011. *Narrative Therapy*. American Psychological Association.

Renee C. B. Manworren and Jennifer Stinson. 2016. Pediatric pain measurement, assessment, and evaluation. In *Seminars in Pediatric Neurology*, Vol. 23. Elsevier, 189–200.

Roger Marek, Cornelia Strobel, Timothy W. Bredy, and Pankaj Sah. 2013. The amygdala and medial prefrontal cortex: Partners in the fear circuit. *Journal of Physiology* 591, 10 (2013), 2381–2391.

David Marino, Paul Bucci, Oliver S. Schneider, and Karon E. MacLean. 2017. Voodle: Vocal doodling to sketch affective robot motion. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, New York, NY, 753–765.

Derek L. Milne and Robert P. Reiser. 2017. *A Manual for Evidence-Based CBT Supervision*. John Wiley & Sons.

Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion Review* 5, 2 (2013), 119–124.

Alan Michael Nevill and Andrew Lane. 2007. Why self-report "Likert" scale data should not be log-transformed. *Journal of Sports Sciences* 25, 1 (2007), 1–2. https://doi.org/10.1080/02640410601111183

K. S. Ong and R. A. Seymour. 2004. Pain measurement in humans. *Surgeon* 2, 1 (2004), 15–27.

André Parent and Lili-Naz Hazrati. 1995. Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Research Reviews* 20, 1 (1995), 91–127.

Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. 2019. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 687–694.

J. Pimentel and J. L. Pimentel. 2019. Some biases in Likert scaling usage and its correction. *International Journal of Science: Basic and Applied Research* 45, 1 (2019), 183–191.

Blaine A. Price, Ryan Kelly, Vikram Mehta, Ciaran McCormick, Hanad Ahmed, and Oliver Pearce. 2018. Feel my pain: Design and evaluation of Painpad, a tangible device for supporting inpatient self-logging of pain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 169.

Sheela Raja, Memoona Hasnain, Michelle Hoersch, Stephanie Gove-Yin, and Chelsea Rajagopalan. 2015. Trauma informed care in medicine. *Family & Community Health* 38, 3 (2015), 216–226.

Jonathan D. Raskin. 2002. Constructivism in psychology: Personal construct psychology, radical constructivism, and social constructionism. *American Communication Journal* 5, 3 (2002), 1–25.

Mark Risjord. 2011. *Nursing Knowledge: Science, Practice, and Philosophy*. John Wiley & Sons.

Elisa M. Rosier, Michael J. Iadarola, and Robert C. Coghill. 2002. Reproducibility of pain measurement and pain perception. *Pain* 98, 1–2 (2002), 205–216.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review* 110, 1 (2003), 145.

James A. Russell. 2015. The greater constructionist project for emotion. In *The Psychological Construction of Emotion*, L. F. Barrett and J. A. Russell (Eds.). Guilford Press, 429–447.

James A. Russell, Anna Weiss, and Gerald A. Mendelsohn. 1989. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology* 57, 3 (1989), 493.

Anke Samulowitz, Ida Gremyr, Erik Eriksson, and Gunnel Hensing. 2018. "Brave men" and "emotional women": A theory-guided literature review on gender bias in health care and gendered norms towards patients with chronic pain. *Pain Research and Management* 2018 (2018), 6358624.

Robert M. Sapolsky. 2003. Stress and plasticity in the limbic system. *Neurochemical Research* 28, 11 (2003), 1735–1742.

Stanley Schachter and Jerome Singer. 1962. Cognitive, social, and physiological determinants of emotional state. *Psychological Review* 69, 5 (1962), 379.

Michael Schwarz. 2009. Is psychology based on a methodological error? *Integrative Psychological and Behavioral Science* 43, 3 (2009), 185–213.

Patrick E. Shrout, Gertraud Stadler, Sean P. Lane, M. Joy McClure, Grace L. Jackson, Frederick D. Clavél, Masumi Iida, Marci E. J. Gleason, Joy H. Xu, and Niall Bolger. 2018. Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences* 115, 1 (2018), E15–E23.

Jeffry A. Simpson. 2007. Foundations of interpersonal trust. In *Social Psychology: Handbook of Basic Principles* (2nd ed.), Arie W. Kruglanski and E. Tory Higgins (Eds.). Guilford Press, 587–607.

Jan E. Stets and Jonathan H. Turner. 2014. *Handbook of the Sociology of Emotions*, Vol. 2. Springer.

TICIRC. 2020. *What Is Trauma-Informed Care? Trauma-Informed Care Implementation Resource Center*. Retrieved June 14, 2022 from https://www.traumainformedcare.chcs.org/what-is-trauma-informed-care/.

Jonathan H. Turner and Jan E. Stets. 2006. Sociological theories of human emotions. *Annual Review of Sociology* 32 (2006), 25–52.

Bessel A. Van der Kolk. 1994. The body keeps the score: Memory and the evolving psychobiology of posttraumatic stress. *Harvard Review of Psychiatry* 1, 5 (1994), 253–265.

Bessel A. Van der Kolk. 2015. *The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma*. Penguin Books.

David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 54, 6 (1988), 1063.

Luke Jai Wood, Ben Robins, Gabriella Lakatos, Dag Sverre Syrdal, Abolfazl Zaraki, and Kerstin Dautenhahn. 2019. Developing a protocol and experimental setup for using a humanoid robot to assist children with autism to develop visual perspective taking skills. *Paladyn, Journal of Behavioral Robotics* 10, 1 (2019), 167–179.

Kaya Yilmaz. 2013. Comparison of quantitative and qualitative research traditions: Epistemological, theoretical, and methodological differences. *European Journal of Education* 48, 2 (2013), 311–325.

Received 1 June 2021; revised 31 March 2022; accepted 19 May 2022