

# Real Emotions Don't Stand Still: Toward Ecologically Viable Representation of Affective Interaction

Paul H. Bucci, X. Laura Cang, Hailey Mah, Laura Rodgers, Karon E. MacLean\*

Computer Science

University of British Columbia

Vancouver, Canada

{pbucci, cang, haileylm, lrodders, maclean}@cs.ubc.ca

**Abstract**—To create emotionally expressive robots, designers of human-robot interaction routinely translate emotion theories into instruments through which we estimate, quantify and analyze human emotional responses to robot behaviour.

Pragmatically, we often use straightforward models such as Russell's circumplex, treating emotion as a single point in a two-dimensional space. However, this simple metaphor and its consequent representations omit many aspects of real emotional experience, can lead to erroneous data and may undermine computational models that rely on them. Problems with emotion representations currently prevalent in human-robot interaction fall into three categories: (1) Representations are static and singular, whereas real emotions can be dynamic, multi-valued, uncertain or conflicting. (2) The framing of an interaction is unspecified (i.e., in an affective rating task: which part of an interaction involving multiple parties and perspectives the participant is meant to consider). (3) Participant responses captured with instruments and methods that are not well-understood by experimenters nor participants produce data that is hard to interpret. We propose alternative emotion representations to account for dynamic emotions inherent in interactive contexts; scrutinize framing ambiguities in study tasks and argue for mixed-methods approaches to achieve shared understanding of emotion representations between participants and researchers.

**Index Terms**—affective computing, human-robot interaction, robot learning, methodology, self-reports

## I. INTRODUCTION

An objective of affective interaction is to create machines that can emotionally interact with humans in real time. In human-robot interaction (HRI), roboticists often draw on emotion theory to evaluate human affect and build computational models that relate human behaviour and biophysical signals to robot behaviours, or vice-versa. This process often takes the form of assigning emotion ratings to robot behaviour, identifying behaviour features, then seeking correlations between these features and the emotion ratings.

Real-time robot behaviour can be generated through a feedback control loop [36] that includes a computational model of human emotion requiring direct behaviour labelling. This loop implies a schema in which the system reasons about the human's emotion, then produces a behaviour which is expected to be an appropriate response to that human's emotion state.

\*All authors contributed equally. Funding from the Natural Sciences and Engineering Research Council of Canada.

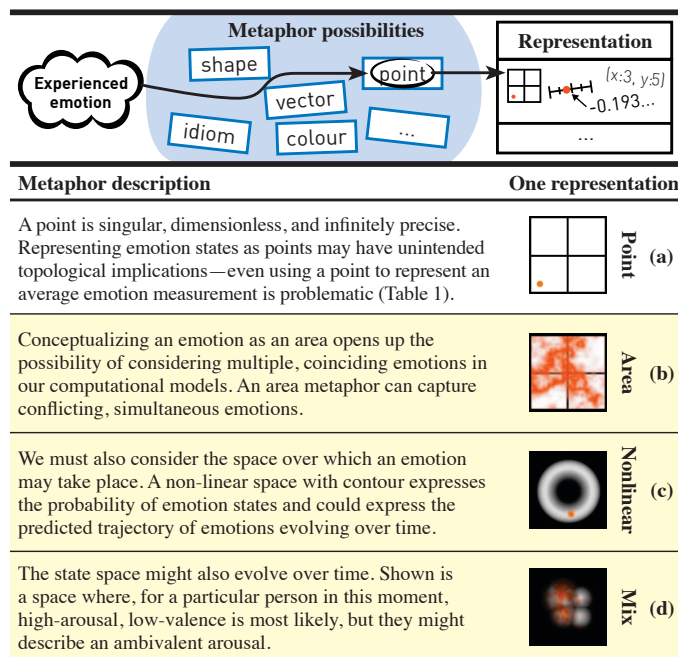


Fig. 1. Experienced emotions can be reasoned about through the use of metaphors: abstract concepts (mathematical, literary, etc.) that stand in for real-world phenomena. Metaphors can be turned into a multitude of concrete representations to serve different purposes. A common metaphor for emotion is a point, which can be represented as a dot on a graph, a decimal, or coordinates. We propose area and non-linear metaphors as alternatives, which enable different ways of conceptualizing emotional experience (yellow).

However, consider human-human emotional interaction in the real world: we need not name another's emotion in order to react emotionally. On the contrary, it often takes significant cognitive effort, perhaps even formal training, to both hold back our reactive instinct and articulate our emotions.

In this position paper, we advance three critiques of HRI studies that rely on emotion labelling, drawing from our own research efforts. By reconsidering how we use common emotion metaphors and representations, frame behaviour labelling tasks, and negotiate meaning in our methodologies, we can get closer to the goal of designing interactive entities whose *behaviour* reflects how we have specified that they should *feel*.

We contribute these problems for the field to consider:

### I. Common metaphors do not account for dynamic emotions.

Representing emotions that change over time, are uncertain, or are in conflict requires amending our current metaphors and representations of emotion.

### II. Contemporary practices do not always explain whose emotion is being measured.

Interaction framing is often unspecified, leaving uncertainty in what an emotion is being ascribed to: a robot's behaviour, a participant's response to the behaviour, or something else.

### III. The meanings of measurement scales are ambiguous.

We often fail to create a shared understanding of measurement scales between participants and researchers.

## II. DEFINITIONS AND APPROACH

To preface our critique, we outline our definitions for metaphors, representations, framing, and shared meaning-making. We then look at how HRI researchers currently use emotion theory to inform their work, produce study instruments, and build computational models.

1) *Metaphors and Representations*: The words “metaphor” and “representation” are sometimes used interchangeably to mean “ideas that stand in for other ideas,” but for the present purpose we require their nuanced distinction.

**Metaphors** can describe phenomena that are otherwise hard to articulate or understand, allowing us to reason and communicate about abstract concepts [19]. For example, saying you have a “white-hot rage” vs. a “simmering rage” relates temperature to emotion, enabling the comparison of emotions via the concept of temperature. Similarly, when we represent an emotion as a single point in a dimensional space, we are using the spatial metaphor of a scalar quantity to communicate differences in an experienced emotion.

To engineer emotional human-robot interactions, we translate our metaphors into concrete **representations** using ink, code, or bits. These representations become the instruments in our studies, shape the input to our algorithms, and contribute directly to our computational models. It is important to clarify the connection between our metaphors and which aspects of emotional experiences they are meant to represent (Figure 1).

Researchers often create metaphors as stand-ins for phenomena, then operationalize the metaphors in order to make predictions: “[*depicting a concept*] as an entity allows us to refer to it, quantify it, identify a particular aspect of it, see it as a cause, act with respect to it, and perhaps even believe that we understand it.” (Lakoff & Johnson [19]).

One representation of the aforementioned metaphor of affect as a scalar quality is Russell's circumplex: an orthogonal space with dimensions of valence and arousal (Figure 1, top right) [25]. While not meant as a direct representation of brain and body, it is useful to think about the human experience of affect as mapping to this space [2]. For example, to communicate with participants about their emotion, we can employ instruments such as the Affect Grid (a discretized 2D circumplex) [26] or the Self-Assessment Manikin (SAM), which splits the arousal-valence-dominance space into three scales with cartoons for each scale item [5].

Our purpose in this detailed inspection of metaphors and their corresponding representations is to better understand both the underlying emotional phenomena and how to operationalize metaphors as representations in computational models.

2) *Emotion Models*: In interactive emotion modeling, this term has multiple uses.

**As an emotion theory**: Models typically instantiate a theory. However, theoretical definitions of models *explain* emotion, e.g., that an emotion exists, that a subjective state is expressible through certain externally-detectable human behaviours, or that emotions can be defined in terms of valence and arousal.

**As a computational model**: A computational model's purpose is to predict human expression and possibly drive system responses, rather than explain them – e.g., a machine learning or artificially intelligent representation used to detect and classify emotions.

**As an instrument**: The tools used for *measuring* emotion in a research context act as a medium of communication between participants and researchers (e.g., the SAM or Affect Grid).

3) *Methodology: framing and meaning-making*: Our approaches to designing, running, analyzing and reporting on our studies greatly influence our computational models and robot control architectures. There is a close link between the social construction of meaning and the practical construction of our real, physical, embodied interactive systems. The way in which we elicit emotion ratings from participants is an integral part of the resulting computational model.

As an example, imagine a study where a participant watches an industrial robot arm perform a series of short pick-and-place tasks. Each participant is given the same written instructions to assess the valence of the robot from stressed–excited on a semantic differential scale. Although the experimenter can answer clarifying questions, current practices encourage them to respond minimally lest they influence the trial.

Some participants imagine that the robot is a persistent conscious entity that is aware of them the whole time. Others imagine that the robot resets its memory between trials. Imagining the former, a participant might see subsequent trials as the robot trying and failing to communicate with them, rating the robot “stressed.” However, this difference in framing would not be captured with a rating scale alone.

In controlled scientific process, we design studies to maximize consistency so we can attribute causality to manipulated variables, reduce bias and improve objectivity/generalizability. However, in the example above, the experimenter cannot know what is actually being measured with the participant ratings, and may not even realize the experiment's potential for ambiguity. The rigor gained by controlling this experiment's conditions is substantially undermined.

Ironically, such error can be a direct consequence of intended rigor: e.g., the concern that experimenter interaction with a participant may actually introduce response bias. At other times, it may be due to belief that a scale's “validation” means it can be deployed without explanation or instruction. In fact, participants may not truly understand what they are

intended to respond/evaluate when given a survey instrument. There are two important methodological considerations here:

By **framing** a study task, we mean articulating what an emotion rating is being ascribed to within that task's context. A participant needs to understand what they are supposed to rate, e.g., how *they* feel, how they imagine a *robot* might feel, or how a robot is *trying to make them* feel (Table II). This is not always an easy distinction to make, nor to instruct.

**Shared meaning-making** refers to a process of resolving ambiguities through discussion between researchers and participants. A failure to do so puts in question understanding both of the interaction tasks, and of response instruments (e.g., rating scales). With the addition of qualitative methods, however, nuances in subjective experience can be addressed.

A first step for the field would simply be a widely accepted realization that the potential for ambiguity exists; and a second, to ensure that qualitative methods (even as basic as an interview) are accepted and required as a standard for both generating and interpreting quantitative data.

### III. RELATED WORK

Recent theoretical work in emotional interaction has challenged the dominant “signalling paradigm” [18] of emotion classification which assumes (1) all relevant information about an interaction is encoded in a signal and (2) there is a universal congruence between social meaning, behaviour, and subjective experience [18], [20]. In our own work, participants have regularly disproven our expectations that study tasks are universally understood, and that study instruments can fully capture how participants feel during an interaction.

It seems common research methodologies and conceptions of emotion measurements that were initially helpful may obfuscate the path forward. Here, we unpack the problems.

**Problem 1:** *Prevalent emotion representations imply that each robot or human behaviour should map to a single emotion regardless of context.*

Researchers in HRI and psychology have begun to recognize that behaviours have context-dependent meaning, which confounds methods that label behaviours with singular emotions [1], [10], [16], [18]. Jung introduces the concept of *affective grounding* to explain how the same signals (e.g., facial expressions, gestures) can vary in emotional and social meaning based on context. An affectively-grounded interaction is one where a signal's meaning is converged upon as a result of continuous interaction (or “emotion coordination”) [18]. However, this perspective is new to the field: reviewing 27 robot expression papers, Fischer *et al.* found the dominant assumption to be that a behaviour can convey an emotion independent of context [13].

The behaviour labelling approach is eminently reasonable: computational models need explicit labels for training data. Dimensional and categorical emotion theories are used to produce self-report instruments that capture participants' emotion ratings of both their own and robot behaviours. Studies use Ekman's theory of basic emotions [8], [12], [13], [18], Russell's dimensional model of affect [3], [9], [23], [27], [30]

or a combination of both [28], [36]. Instruments include the Affect Grid [26], the Self-Assessment Manikin [21], [27], or the PANAS scales [1].

Herein lies the dilemma: computational models of behaviour require labels, but behaviours cannot be consistently and directly labeled with a single emotion [20]. We could add contextual details to computational models to improve labelling accuracy [6], [10], [11]. Alternatively, we could actively choose to represent conflicting or mixed emotions, aligning more closely with how behaviours are experienced and interpreted in real life [9]. We present a discussion of alternative representations in Section IV.

**Problem 2:** *Experimental paradigms overlook pervasive framing ambiguities in rating emotions during interactions.*

Framing a human-robot interaction task is like directing a participant to empathize: participants can be asked to either *recognize* or *experience/respond* to emotional robot behaviours [15]. Failing to specify which is called for can result in a participant misunderstanding their job and generating data irrelevant to the experimental intent (a situation we experienced in our own work).

Meanwhile, many HRI articles do not specify either instructions or intent, leaving readers uncertain what the results mean. As an example: we examined the 52 full, peer-reviewed papers published in the HRI'18 conference [17]. 26 reported studies where participants judged affect. Of these, in 9, task framing was clear to readers and participants. In 3, framing was clear only in some respects. In 14, it was substantially ambiguous. We offer [29], [32], [35] as excellent framing examples. Robots are introduced as situated in the task, participants can conceptualize the interaction prior to rating, and experimenters listen to and iterate with participants to establish meaning.

Fortunately, there are ways to avoid this situation without evident compromise of scientific rigor. Some HRI studies implicitly explicate frame by asking contrasting questions using different frames [8], [9], [23]. Others establish frame through clarifying interviews where participants explain their interpretation of the study task [10], [20]. Still others use concepts from theatre. Bucci *et al.* establish roles, characters, and settings for an interactive scene [10]. Westlund *et al.* do this through an interactive theatrical process [34]: participants (children) are introduced to a puppet who has a strong personality, a reason for being there, and a name. The puppet then introduces the robot to the participants, clearly addressing the relationship between all actors. Marino *et al.* offered improvisation as a way for participants to design robot emotion-transition behaviours, who found the design tasks easier once an interaction was framed in a scene [22].

In summary, we can see multiple ways of establishing the frame of a study task so as to direct a participant's effort to the kind of empathy the researcher wants to inspect.

**Problem 3:** *Experimental paradigms rely on participants and researchers having a mutual understanding of study instruments that measure universal quantities of emotion.*

Self-report instruments such as Likert scales and the Affect

TABLE I

Dimensional theories of emotion use the metaphor of multi-dimensional scalar quantities to reason about subjective experiences. Because our metaphors will be represented in computer code, we must use metaphors more literally than they may have been intended. Here we outline the implicit assumptions and consequences of strictly interpreting emotions as a point on a linear, dimensional space. This table elaborates on *Problem 1* from *Related Work*.

<b>Implicit Assumption 1: Emotions can be represented as a single point-like state</b>			
<i>Implication of making assumption</i>	<i>Ensuing representation limitation</i>	<i>Example of experience mismatch</i>	<i>Representation/experience mismatch</i>
<i>Focus</i> : One’s emotional state must be identified as a singular, focused point in space.	A single point does not allow for the representation of multiple, conflicting emotions.	I am happy I got a new job but am also nervous at the same time. How do I represent this feeling as a point?	An emotion is not always experienced singularly: they can be conflicting, mixed, or multiple.
<i>Fixedness</i> : Over a period of time, one can experience only a single fixed emotion, which cannot change.	Experiencing emotion does not feel like a series of single moments: rather, it is dynamic and appears to continuously change.	During a task, I am surprised briefly but otherwise neutral. How do I describe my emotional state over the entire period of time?	Asking for a single point to represent an emotional experience hides the variation people feel over time during the experience.
<b>Implicit Assumption 2: Emotion space is continuous and linear</b>			
<i>Implication of making assumption</i>	<i>Ensuing representation limitation</i>	<i>Example of experience mismatch</i>	<i>Representation/experience mismatch</i>
<i>Linearity</i> : Emotions must be distinct within the space; linear, equidistant points correspond to similar magnitudes of emotion differences.	It may be difficult to convey the magnitude of qualitative differences in felt emotions by identifying discrete points on a line.	It takes more effort for me to become extremely happy than a little bit happy. How do I indicate the magnitude of effort?	By default, emotion rating scales are linear and uniform. However, not all perceptions are linear (e.g., perceptions of loudness are exponential).
<i>Probability</i> : Each point must be as accessible or likely to be reached as all others.	A flat, unweighted space does not express that some emotions are more difficult to feel and may be dependent on previous emotions.	If I’m feeling good when someone snaps at me, I’m less likely to feel angry than if I was already upset. How do I express this likelihood?	Some emotions are more unlikely or more difficult to experience, (e.g., extremes or true neutrals).
<i>Unclear Temporality</i> : If the space is projected into time, instantaneous transitions between extreme emotion states are not allowed.	Traversal from one emotional state to another can feel instantaneous, as well as discontinuous; and transitions are not the same every time.	I feel like I can transition from happy to angry without passing through a neutral-valence state.	The 2D Affect Grid gives no guidance on which emotion transitions are natural—how do you move from place to place?

Grid usefully allow a participant to report quantitatively on their own subjective experiences. However, people naturally differ in interpreting a scale’s “distances” relative to the emotional quantity it represents [33]. There are examples of scales measuring subjective, affect-related quantities, such as pain, where research has found that baseline and extrema depend on personal experience (e.g., the worst pain you have ever felt is different than mine). Accepted practice with pain scales recognizes that meaning can be relative to a treatment program, and may need significant discussion to situate the scale in the rater’s personal history of pain [7], [24], [31].

Our own experience of scales like the Affect Grid has exposed variance in user understanding of scale meaning. Their first impressions may not correspond to what experimenters expect to measure, e.g., with respect to scale linearity.

HRI researchers have been arguing for stronger integration of qualitative and quantitative research designs (“mixed-methods”) that include participants directly in the co-construction of meaning: collaboratively understanding the rating scales [4], [14], [18]. Co-constructing means that experimenters can define the structure of the scale (e.g., one-dimensional, 5-item, linearity, etc.), and allow participants to explicate the scale boundaries relative to the specified interactive context and participant’s own experience. The resulting relative scale enables clearer between-participant comparison without presuming that a subjective experience has some absolute, objective quantity.

Leahu and Sengers emphasize working with participants to define what emotion words mean. They “expose the [computational] models” by reviewing qualitative/quantitative results

together with participants; we further emphasize that scale calibration needs to happen *prior* to use of the scale even if post-hoc review is needed. We present a process for a mixed-methods approach to defining the meaning of study instruments between participants and experimenters in Section VI.

**Takeaways:** Interactive affect research has reached a state where: (1) We require representations of emotion that can convey uncertainty, motion and mixing. (2) Study tasks are rarely framed explicitly, but there are examples of doing this without impacting experimental rigor. (3) Study instruments and methods, even when validated, can be interpreted individually, undermining accuracy; one safeguard is a method whereby experimenters work with participants to personally relate their experience to the provided scale within the interaction context.

In the following, we expand on our arguments and make concrete recommendations for the field to consider.

#### IV. MODEL METAPHORS

Building computational models of affect requires collecting quantitative emotion data or labels. The instruments we choose for measuring this data are a product of the metaphors we use to describe and explain the emotional experience. Selecting a metaphor appropriately has the power to communicate the researchers’ interpretation of the emotion space, and consequently align participants to the same understanding.

Dimensional theories of affect and communication use the metaphor of multi-dimensional scalar qualities to reason about subjective experience. Here, we articulate and critique two assumptions (Table I) about the emotion space implicit in these metaphors: (1) that emotions can be represented as a single

point-like state, and (2) emotion space can be conceptualized as continuous and linear. These assumptions structure both how emotions can be conceptualized and how emotions can be represented using instruments within an experimental context.

First, the common usage of a point-like metaphor for emotions implies that one's current emotional state can be unambiguously captured for a given instant. However, in real-life emotional interactions, our experience is rarely focused to a single point: as events play out, we evolve our own understanding of emotions as well as our evaluations of others' [2]. We might also experience multiple or conflicting emotions.

Second, the common circumplex representation implies a topology in which the space can be traversed consistently, with equal probability of reaching the entire space. Yet, movement between emotion states is not so tidy; there is more to represent than a linear movement through a uniform orthogonal space. Does a continuous space represent all possible emotions a person could feel? If each point in the space represents an emotion state, then does inhabiting different points in the space feel different? Do we experience emotions independently? To address the first assumption, we propose alternative metaphors for the unit of representation for emotional states. For the second, we suggest different emotion space topologies.

#### A. Area metaphors: representing emotion state

Asking participants to identify an emotion as a point in a space implies that they are *capable* of identifying the emotion, they are experiencing only one, and their experience is static. Consider an alternative metaphor: think of the emotion representation as an *area* to better encompass the real-life complexity of mixed, conflicting and dynamic emotions in ourselves, or uncertainty in attributing emotion to an agent's behaviours.

Emotions evolve in an interactive context. This *temporal* aspect necessitates that we use more than a single point to represent emotion states over time. An area metaphor can capture movement through the emotion space over time, as illustrated in Figure 1. We claim that uncertainty should be directly accounted for in any representation, not simply as error, but as fundamental to what it means to experience emotions ourselves and ascribe it to behaviours. Researchers often analyze robot behaviour in terms of averages of Likert scale measurements. Using the average implies there is a precise point-like emotion that a particular robot behaviour *should* convey, and that deviations from that theoretical average are measurement errors. Remove the concept of a point-like emotion, and it becomes reasonable to talk about the behaviour's inhabiting a probability distribution over an emotion *space*, where this space itself represents the possibility of the emotion the behaviour may connote. A behaviour may not convey the same emotion each time (it is not deterministic); our representations should account for this.

#### B. Nonlinear spaces: topography of possible emotion states

The metaphorical emotion space should also represent the possible emotions that a person can feel. Descriptively, there

are portions of the emotion space that are more difficult to attain, e.g., it is more rare and perhaps effortful to be ecstatic than to be depressed. Imbuing the emotion space itself with contour allows for representations of a directional quality or likelihood of moving from one emotion to another (see (c) and (d) in Figure 1 for examples of contoured emotion spaces).

In modeling interactive emotions, we might think of the space itself changing over time: as you feel more sad, it might be easier to get angry than relaxed, despite these being separated by similar Euclidean distances on the Affect Grid. In such a case, an emotion experience is not simply a *point* but a *trajectory* over a perpetually reforming terrain.

#### C. Alternative Representations

We present the above alternative representations to challenge the norm and widen the space of metaphors we currently use. We invite fellow researchers to consider the implicit metaphorical claims of their chosen representations when designing studies, and ground them in their participants' subjective experiences. As researchers who build interactive emotion models, we posit that **representations** should feature:

**RF1. Multiple points**, due to the human experience of conflicting emotions.

**RF2. Model uncertainty estimates**, reflecting ambiguity in how we experience emotion.

**RF3. Time-variance**, for movement through emotion space.

**RF4. Non-linearity**, with collection instruments that support responses that move on different topologies.

## V. FRAMING PROBLEMS






Picture a slapstick comedian performing a banana-peel bit in front of a live audience. The comedian trips, falls loudly and screws up their face in pain. The audience laughs. We could ask the audience, "How did this performance make you feel?" or "What feeling is the comedian expressing during this act?". The ratings would differ wildly depending on what the audience thought the framing of the rating task was, as each has a different meaning [18].

In an interaction rating task, there is an evaluator and something that is being evaluated. There is ambiguity in whether a participant is meant to evaluate how they feel, or to guess what another thing is supposed to feel. As illustrated in Table V, there are a number of possible **framings** between one participant and one robot, each of which would attribute an emotion rating to a different aspect of an interaction. The methods we use should disambiguate these framings to ensure the reliability of gathered data.

Many of the instruments we employ were originally designed for self-report of one's own affective state. For example, the SAM is intended as an easily understood, culturally universal method for a participant to express their internal affect via cartoon depictions of the body [5]. When rating a robot's behaviour with the SAM, the implicit assumption of the experimental task could be that: (1) the behaviour makes a participant feel an emotion; (2) the robot's behaviour

TABLE II

During an experiment, it is sometimes unclear which portion of an emotional interaction we are asking participants to consider. Here are possible frames of reference that an experiment could be inspecting.

Cartoon	Description
	<b>Participant (Jan, left) is evaluating how she feels about Robot (Can, right).</b> Jan is being asked to interpret her subjective feelings about how Can is making her feel.
	<b>Jan is evaluating what Can is trying to convey.</b> Jan is being asked to interpret Can's communicative behaviour. Can's expressions give <i>evidence</i> for a hidden subjective state.
	<b>Jan is evaluating how Can feels.</b> Jan is being asked to interpret a set of behaviours over some duration that indicate Can's emotional state.
	<b>Jan is evaluating how Can feels about her.</b> Jan is asked to evaluate how Can is evaluating her subjective state. Jan might view Can's actions to do this, or might consider her own actions.
	<b>Jan is evaluating how she currently feels.</b> Jan is being asked to inspect her body/brain and describe some kind of mixture of mood, emotion, affect, or physiological perceptions.

consistently conveys an emotion; (3) or the robot feels an emotion. The participant may not share the assumption of the experiment with the researcher, nor the understanding that the SAM instrument is intended to be self-reflexive.

In robot emotion studies, directives to rate “the robot’s behaviour,” or even “how the robot feels” are ambiguous. Feeding the resultant corrupt data into a computational model will produce erroneous results. Rather than assume that the intent behind a rating question is obvious to the participant, we suggest that the researcher should:

- F1. Resolve the frame** through calibration via participant discussion or attention to scene-setting.
- F2. Report the framing process** when sharing results, so others can assess their validity and build on them.

## VI. AN ARGUMENT FOR MIXED-METHODS EVALUATION

While the goal of an interactive emotion study is often a quantitative measurement, methods and instruments must use language or images as descriptors to convey meaning. The interpretations of these descriptors vary between people due to their different experiences in the world, which exposes an inherent qualitative aspect in a seemingly quantitative measurement. We suggest embracing this fundamental “mixedness” by ensuring that the meanings of descriptors are well established.

Embracing mixed-methods approaches in our experimental design necessitates: (1) grounding participants in the premise of the interaction; (2) creating shared understanding of instruments and measured phenomena; and (3) creating closer

alignment between experiments and possible real-world applications. Conversation between participants and researchers is required to ground the framing and meaning of study materials and activities. The goal is to *calibrate* participants on the researchers’ intended parameters, but also to *capture* the participants’ experiential richness that has led to their rating.

Specifically, we suggest actively collaborating with participants to ground emotion measurement in personal experience to align quantitative representation and qualitative meaning. Researchers should provide the instrument structure (e.g., the intended subjective spacing between scale elements) and work with participants to explicate the semantic difference of scale items. Researchers should also iteratively assist participants in attributing their experiences to scale items, taking care to ensure that both parties can reason about and refer to the scale similarly. A calibration process allows researchers to assess agreement between participants and report on the accessible emotion range of the interaction. This will generally require the researcher to use a **methodology** in which they:

- M1. Establish the extrema of a scale** by asking a participant to recount events in the interaction.
- M2. Establish the meaning of subjective distance between items** by asking a participant to explain their understanding of each item.
- M3. Converge on researcher-provided structure** by iterating on the above before the scale is used or if meaning shifts during scale use.

Rather than leaving participants’ interpretation of task framing and instruments ambiguous, such a process acknowledges and addresses variation. By explicating the meaning of what is being measured, ambiguities around framing and instrument meaning can be accounted for and, ideally, resolved.

## VII. CONCLUSION

In this paper, we discuss challenges in representing and capturing emotions during interactive emotion studies. We articulate emotion metaphors and representations in common use which shape how emotional experiences are understood, and have a cascading effect on how we collect, analyze and discuss emotional interaction data. Current metaphors are representationally limited in not accounting for time variance and the inherent uncertainty in self-reporting emotion. We propose alternative metaphors based on areas or non-linear topologies that align more closely with the semantics of emotion rating tasks. We identify methodological problems: the framing of emotion tasks can be ambiguous, resulting in categorically confused studies. As a solution, we suggest that a mixed-methods approach of incorporating meaning-making into quantitative research designs will ground the meaning of study instruments and resolve framing problems.

## ACKNOWLEDGMENT

This research was enabled by facilities and additional resources provided by UBCs Institute for Computing, Information and Cognitive Systems (ICICS) and Designing for People (DFP) Research Cluster. And thanks to Anushka Agrawal.

## REFERENCES

- [1] Kaveh Bakhtiyari and Hafizah Husain. Fuzzy model of dominance emotions in affective computing. *Neural Computing and Applications*, 25(6):1467–1477, 10 2014.
- [2] Lisa Feldman Barrett and James A Russell. *The psychological construction of emotion*. Guilford Publications, 2014.
- [3] Emily Bhuwalka, Kunal; Icel, Nur; Gong. How does Robot Feedback Affect Participant Affinity and Trust ? *HRI*, 2018.
- [4] Kirsten Boehner, Rogério Rogerio Depaula, Paul Dourish, and Phoebe Sengers. How emotion is made and measured. *Int. J. Human-Computer Studies*, 65:275–291, 2007.
- [5] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.
- [6] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. Crowdsourcing Human-Robot Interaction: New Methods and System Evaluation in a Public Environment. *Journal of Human-Robot Interaction*, 2(1):82–111, 2013.
- [7] Harald Breivik, PC Borchgrevink, SM Allen, LA Rosseland, L Romundstad, EK Breivik Hals, G Kvarstein, and A Stubhaug. Assessment of pain. *BJA: British Journal of Anaesthesia*, 101(1):17–24, 2008.
- [8] Mason Bretan, Guy Hoffman, and Gil Weinberg. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human Computer Studies*, 78:1–16, 2015.
- [9] Paul Bucci, Xi Laura Cang, Anasazi Valair, David Marino, Lucia Tseng, Merel Jung, Jussi Rantala, Oliver S Schneider, and Karon E MacLean. Sketching cuddlebits: coupled prototyping of body and behaviour for an affective robot pet. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3681–3692. ACM, 2017.
- [10] Paul Bucci, Lotus Zhang, Xi Laura Cang, and Karon E MacLean. 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*, page 509. ACM, 2018.
- [11] Oliver Damm, Karoline Dreier, Frank Hegel, Petra Jaecks, Prisca Steneken, Britta Wrede, and Martina Hielscher-Fastabend. Communicating emotions in robotics: Towards a model of emotional alignment. In *Proceedings of the workshop Expectations in intuitive interaction on the 6th HRI International conference on Human-Robot Interaction*, 2011.
- [12] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.
- [13] Kerstin Fischer, Lars Christian Jensen, Maria Vanessa, and Der Wischen. Emotion Expression in HRI When and Why. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 29–38, 2019.
- [14] Ge Gao, Malte F Jung, Gabriel Culbertson, Susan R Fussell, Malte F Jung, Sun Young Hwang, Gabriel Culbertson, Susan R Fussell, and Malte F Jung. Beyond Information Content: The Effects of Culture On Affective Grounding in Instant Messaging Conversations. *Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article*, 1(18):1–18, 2017.
- [15] JR Hodges. Making it up and making do: Simulation, imagination, and empathic accuracy. *The Handbook of Imagination and Mental Simulation*, pages 281–294, 2008.
- [16] Tom Hollenstein. State space grids. In *State Space Grids*, pages 11–33. Springer, 2013.
- [17] International Conference on Human-Robot Interaction. *HRI '18: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 2018. ACM.
- [18] Malte F Jung. Affective Grounding in Human-Robot Interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pages 263–273, 2017.
- [19] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.
- [20] Lucian Leahu and Phoebe Sengers. Freaky: performing hybrid human-machine emotion. *Designing Interactive Systems*, pages 607–616, 2014.
- [21] Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1), 2018.
- [22] David Marino, Paul Bucci, Oliver S Schneider, and Karon E MacLean. Voodle: Vocal doodling to sketch affective robot motion. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 753–765. ACM, 2017.
- [23] Toru Nakata, Tomomasa Sato, and Taketoshi Mori. Expression of emotion and intention by robot body movement. *5th Conference on Intelligent Autonomous Systems*, pages 352 – 359, 1998.
- [24] Blaine A Price, Ryan Kelly, Vikram Mehta, Ciaran McCormick, Hanad Ahmed, and Oliver Pearce. 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*, page 169. ACM, 2018.
- [25] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980.
- [26] James A Russell, Anna Weiss, and Gerald A Mendelsohn. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology*, 57(3):493, 1989.
- [27] Martin Saerbeck and Christoph Bartneck. Perception of affect elicited by robot motion. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10*, page 53, 2010.
- [28] Jelle Saldien, Kristof Goris, Bram Vanderborcht, Johan Vanderfaellie, and Dirk Lefeber. Expressing emotions with the social robot probo. *International Journal of Social Robotics*, 2(4):377–389, 2010.
- [29] Solace Shen, Petr Slovak, and Malte F Jung. Stop. i see a conflict happening.: A robot mediator for young children’s interpersonal conflict resolution. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 69–77. ACM, 2018.
- [30] Sichao Song and Seiji Yamada. Expressing emotions through color, sound, and vibration with an appearance-constrained social robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 2–11. ACM, 2017.
- [31] Jennifer N Stinson. Improving the assessment of pediatric chronic pain: harnessing the potential of electronic diaries. *Pain Research and Management*, 14(1):59–64, 2009.
- [32] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scasellati. The ripple effects of vulnerability: The effects of a robot’s vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 178–186. ACM, 2018.
- [33] Gail M Sullivan and Anthony R Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013.
- [34] Kory Westlund, M Jacqueline, Sooyeon Jeong, Hae W Park, Samuel Ronfard, Aradhana Adhikari, Paul L Harris, David DeSteno, and Cynthia L Breazeal. Flat vs. expressive storytelling: Young childrens learning and retention of a social robots narrative. *Frontiers in human neuroscience*, 11:295, 2017.
- [35] Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. Thank you for sharing that interesting fact!: Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 298–306. ACM, 2018.
- [36] S. Yohanan and K. E. MacLean. Design and assessment of the haptic creatures affect display. In *ACM/IEEE Intl Conf on Human-Robot Interaction (HRI 11)*, pages 473–480, Lausanne, SW, 2011.