

From Devices to Data and Back Again:
A Tale of Computationally Modelling Affective Touch

by

Xi Laura Cang

BSc Mathematics, University of British Columbia, 2007

BEd Mathematics, University of British Columbia, 2010

B. Computer Science, University of British Columbia, 2014

MSc Computer Science, University of British Columbia, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES

(Computer Science)

The University of British Columbia
(Vancouver)

April 2024

© Xi Laura Cang, 2024

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

**From Devices to Data and Back Again:
A Tale of Computationally Modelling Affective Touch**

submitted by **Xi Laura Cang** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Science**.

Examining Committee:

Karon E. MacLean, Professor, Computer Science, UBC
Supervisor

Rebecca Todd, Associate Professor, Psychology, UBC
Supervisory Committee Member

Cristina Conati, Professor, Computer Science, UBC
Supervisory Committee Member

Todd Handy, Professor, Psychology, UBC
University Examiner

Sid Fels, Professor, Electrical and Computer Engineering, UBC
University Examiner

Nadia Bianchi-Berthouze, Professor, Psychology and Language Sciences, University College London
External Examiner

Abstract

Emotionally responsive Human-Robot Interaction (HRI) has captured our curiosity and imagination in fantastical ways throughout much of modern media. With touch being a valuable yet sorely missed emotion communication channel when in-person interaction is unrealistic for practical reasons, we could look to machine-mediated ways to bridge that distance. In this thesis, we investigate *how* we might enable machines to recognize natural and spontaneous emotional touch expressions in two parts.

First, we take a close look at ways machines engage with human emotion by examining examples of machines in three emotionally communicative roles: as a passive witness receiving and logging the emotional state of their (N=30) human counterparts, as an influential actor whose own breathing behaviour alters human fear response (N=103), and as a conduit for the transmission of emotion expression between human users (N=10 dyads and N=21 individuals).

Next, we argue that in order for devices to be truly emotionally reactive, they should address the time-varying and dynamic nature of emotional lived experience. Any computational or emotion recognition engine intended for use under realistic conditions should acknowledge that emotions will evolve over time. Machine responses may change with changing ‘emotion direction’ – acting in an encouraging way when the user is *happy and getting happier* vs. presenting calming behaviours for *happy but getting anxious*. To that end, we develop a multi-stage emotion self-reporting procedure for collecting N=16 users’ dynamic emotion expression during videogame play. From their keypress force controlling their in-game character, we benchmark individualized recognition performance for emotion direction, even finding it to exceed that of brain activity (as measured by continuous Electroen-

cephalography (EEG)). For a proof-of-concept of a training process that generates models of true and spontaneous emotion expression evolving with the user, we then revise our protocol to be more flexible to naturalistic emotion expression. We build a custom tool to help with data collection and labelling of personal storytelling sessions and evaluate user impressions (N=5 with up to 3 stories each for a total of 10 sessions).

Finally, we conclude with actionable recommendations for advancing the training and machine recognition of naturalistic and dynamic emotion expression.

Lay Summary

This work explores the concept of emotionally responsive Human-Robot Interaction (HRI) and the potential for machines to recognize and interact with natural and spontaneous emotional touch expressions, in two parts. First, we examine three roles through which machines can engage with human emotion: as passive witnesses of human emotion, as influential actors on emotion experiences, and as conduits for emotion expression between human users. Second, we argue that for devices to be truly emotionally reactive, they should address the time-varying and dynamic nature of emotional experience. To this end, we present a multi-stage emotion self-reporting procedure and a proof-of-concept for a training process to generate models of spontaneous emotion expression. To conclude, we reflect on important considerations for designing devices intended to engage with naturalistic and dynamic emotion expression.

Preface

The bulk of this thesis is comprised of papers that are previously published (Ch 3, Ch 5-7), submitted for review (Ch 4), or in preparation for submission (Ch 8) at the time this thesis was completed. In all cases, I have had the great fortune to have learned from and collaborated with colleagues, co-authors, and my PhD supervisor, Prof. Karon E. MacLean. For these chapters, I detail current publication status and my direct contributions to each.

The work in Chapter 3 was completed in partnership with Facebook Reality Labs (now Meta Reality Labs), where the project was initially conceived of during an internship under the supervision of Dr. Ali Israr and later refined together with Prof. MacLean. The protocol was approved by the USA Western Institutional Review Board (WIRB ref# AGHM-2019) and is included here as published in 2023 as “When is a Haptic Message Like an Inside Joke? Digitally Mediated Emotive Communication Builds on Shared History” at IEEE Transactions on Affective Computing [45].

Cang, Xi Laura, Ali Israr, and Karon E. MacLean. “When is a Haptic Message Like an Inside Joke? Digitally Mediated Emotive Communication Builds on Shared History.” *IEEE Transactions on Affective Computing* 14.1 (2023): pp.732-746.

Chapter 4 is the result of a close collaboration with UBC Psychology, where the entire team worked together from project inception to implementation. Dr. Zak Witkower (a PhD student in Psychology at the time) was lead author, managing data collection, the large majority of the final writing and analysis, and was supported by his supervisor Prof. Jessica Tracy (UBC Psychology). As second au-

thor, I contributed to the robot breathing behaviour generation, data recording and formatting software, study design and analysis interpretation as well as writing and reviewing key passages. Other co-authors include Paul Bucci who designed the physical robot, contributing in equal measure on software development and study design; we were both supported throughout by Prof. MacLean. The work is currently under review.

Witkower, Zak, Xi Laura Cang, Paul Bucci, Karon E. MacLean, and Jessica Tracy. “Catching Fear from a Non-Living, Artificially Breathing Organism: Human Psychophysiology is Guided by a Robot Displaying Distinctive Respiratory Patterns.” *In review*.

I led the production of the work in Chapter 5, from study design to touch sensing and data collection software to data analysis and formal writing. Two co-authors were invaluable contributors at all stages, and individually responsible for feature extraction (Paul Bucci for biosignal data, and Dr. Jussi Rantala for gaze data). The paper [42] is presented as published.

Cang, Xi Laura, Paul Bucci, Jussi Rantala, and Karon Maclean. “Discerning affect from touch and gaze during interaction with a robot pet.” *IEEE Transactions on Affective Computing* (2021): pp.1598-1612.

Chapter 2 is also presented as published¹(ACII 2019) – a conference paper titled “Real Emotions Don’t Stand Still: Toward Ecologically Viable Representation of Affective Interaction.” This work [39] was an intellectual exercise which benefited from (at times) intense input of a number of different perspectives by authors in equal contribution across all development phases. From ideation to writing, myself and Paul Bucci (both PhD students) were supported by Dr. MacLean as well as undergraduate students Hailey Mah and Laura Rodgers. Hailey and Laura were also heavily involved in projects presented later in this thesis.

Bucci, Paul H., X. Laura Cang, Hailey Mah, Laura Rodgers, and Karon E. MacLean. “Real emotions don’t stand still: Toward ecologic-

¹With the addition of a few relevant references .

ally viable representation of affective interaction.” In 2019 8th *International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019: pp. 1-7.

Chapter 6 is the first publication from a project that had an exceptionally complex and labor-intensive study and protocol, taking four years to design, collect, and analyze. Over the full life of the project, many graduate and undergraduate researchers took part to different degrees. The author list includes 11 team members who were significantly involved in protocol testing and iterative refinement (Paul Bucci), mixed methods data collection and coordination (Laura Rodgers, Hailey Mah, Anushka Agrawal), exploratory analysis (Rubia Guerra, Bereket Guta, Paul Bucci, Laura Rodgers, Shinmin Hsu, Qianqian Feng, Chuxuan Zhang), software development particularly for custom hardware like the unbiased joystick for labelling and data synchronization (Laura Rodgers), data visualization (Rubia Guerra, Bereket Guta), and writing and editing (Rubia Guerra, Bereket Guta, Qianqian Feng). A number of other experts and volunteers providing valuable support are mentioned in the papers’ acknowledgements list. As first author, I led the project through all phases and presented the work at the ACII conference.

Cang, Xi Laura, Rubia R. Guerra, Paul Bucci, Bereket Guta, Karon MacLean, Laura Rodgers, Hailey Mah et al. “Choose or Fuse: Enriching Data Views with Multi-label Emotion Dynamics.” In 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2022: pp. 1-8.

I was first author of the published paper featured as Chapter 7, leading all phases of project development with the early-listed authors (myself, Rubia Guerra, and Bereket Guta) contributing most prominently to the featured analysis, generating models, running experiments, and paper writing. Paul Bucci, Hailey Mah, and Laura Rodgers played instrumental roles in early project development and all data collection (as described in Chapter 6). Much of the early analysis inspired the final content in the paper and is thanks to Paul and Hailey’s detail-rich qualitative data coding and Laura Rodgers’ intensive multi-stream data alignment and visualization scripting.

Cang, X. Laura, Rubia R. Guerra, Bereket Guta, Paul Bucci, Laura Rodgers, Hailey Mah, Qianqian Feng, Anushka Agrawal, and Karon E. MacLean. “FEELing (key) Pressed: Implicit Touch Pressure Bests Brain Activity in Modelling Emotion Dynamics in the Space Between Stressed and Relaxed.”, *IEEE Transactions on Haptics* (2023): pp. 1-8.

Content from Chapter 8 is intended for inclusion in a paper submission led by Rubia Reis Guerra, supported by myself (as second author), Daniel Chen, and Nao Rojas, all supervised by Prof. MacLean. I worked very closely with Rubia in planning the study design and executing the data collection. The data collection interface was implemented by Nao Rojas under close consultation and repeat piloting with Rubia and myself. Analysis was a collaborative effort whereby Rubia and Daniel constructed and tabulated classification results with input on experiment structure from myself and Prof. MacLean. I conducted qualitative analysis with input from co-authors. My primary contributions are the focus of Chapter 8 and, while key passages are to appear in a planned journal submission and Rubia’s thesis, there will likely be important variations in later versions.

Unless otherwise specified, the studies in this dissertation were reviewed and approved by the University of British Columbia (UBC) Behavioural Research Ethics Board (BREB) with ethics number #H15-02611.

I am greatly indebted to everyone listed here as well as those named in each projects’ acknowledgements and many others in my community who have given invaluable feedback and served as critical sounding boards to vastly improve this work in all ways. While they have had great influence in the intellectual contributions, I take responsibility for the concept of this thesis, chapter integration, and all other formal writing, including any and all errata.

Contents

Abstract	iii
Lay Summary	v
Preface	vi
Contents	x
List of Tables	xvii
List of Figures	xx
List of Abbreviations	xxvii
Acknowledgments	xxviii
Dedication	xxx
1 Introduction	1
1.1 A Note on Terminology	3
1.1.1 Social, Affiliative and Affective Touch	3
1.1.2 Authentic and Spontaneous Emotion	3
1.2 Reading Emotional Touch	4
1.2.1 As Conduit	6
1.2.2 As Influence	7
1.2.3 As Witness	8

1.3	Prerequisites for Machine Recognition of Emotion in Touch . . .	12
1.3.1	Real Emotion Elicitation	12
1.3.2	Emotion-Embedding Modalities	13
1.3.3	Conceptualization and Labelling	15
1.4	Thesis Contributions and Organization	16
I	Emotion Communication with Machines	19
2	Metaphors for Emotion: An Argument for Rich Emotion Labelling	22
2.1	Introduction	23
2.2	Definitions and Approach	25
2.3	Related Work	28
2.4	Model Metaphors	33
2.4.1	Area metaphors: representing emotion state	34
2.4.2	Nonlinear spaces: topography of possible emotion states .	35
2.4.3	Alternative Representations	35
2.5	Framing problems	36
2.6	An Argument for Mixed-Methods Evaluation	38
2.7	Conclusion	39
3	Machine as Emotion Conduit: An Example of Haptic Messaging in Emotion-Laden Scenarios	40
3.1	Introduction	41
3.2	Background	44
3.3	Materials and Methods	47
3.3.1	The Haptic Display	47
3.3.2	Message Meaning Phase: MTurk Online Survey	53
3.3.3	Design Phase: In-Person Dyads	54
3.3.4	Interpretation Phase: In-Person Singles	57
3.4	Analysis & Results	59
3.4.1	Features and Parameter Analysis	59
3.4.2	Qualitative Analysis of Participant Designs	63
3.4.3	Interpretation Accuracy	64

3.5	Discussion	69
3.5.1	Message Design Observations	69
3.5.2	Interpretation Rate	71
3.5.3	The Messaging Experience	73
3.6	Conclusions	76
4	Machine as Emotion Influence: An Investigation into Machine Breathing as a Fear Contagion	78
4.1	Introduction and Background	79
4.1.1	The Current Research	81
4.2	Methods	82
4.2.1	Participants	82
4.2.2	Procedure	82
4.2.3	Materials	86
4.3	Results	88
4.3.1	Manipulation Checks	88
4.3.2	Main Analyses: Do humans show emotion contagion from artificially breathing robots, via touch?	88
4.4	General Discussion	92
4.4.1	Limitations and Future Directions	93
4.4.2	Advances in Human-Robot Interaction (HRI)	95
5	Machine as Emotion Witness: A Study of Machine Classification of Emotion from Personal Storytelling	96
5.1	Introduction	97
5.1.1	Approach and Research Questions	98
5.1.2	Contributions	102
5.2	Related Work	102
5.2.1	Targeted Emotion Set	103
5.2.2	Elicitation of True Emotion	103
5.2.3	Recognition Modalities	104
5.3	Methods	106
5.3.1	Data Collection	106

5.3.2	Features, Pre-Processing, Extraction & Analysis	110
5.4	Results	117
5.4.1	Subject-Independent Emotion Classification	117
5.4.2	Participant Classification	118
5.4.3	Subject-Dependent Emotion Classification	118
5.4.4	Feature Set Analysis	122
5.4.5	Reports of Experienced Emotion	123
5.5	Discussion	125
5.5.1	RQ1: Ability of Touch and Gaze to Predict Emotion . . .	126
5.5.2	RQ2: Individuality	127
5.5.3	RQ3: Sample Density for Realtime Responsiveness . . .	127
5.5.4	RQ4: Experimental Methodology	129
5.5.5	Implications for Social Robot Applications	130
5.6	Conclusions	131

II Dynamic Emotion Modelling 133

6	Dynamic Emotion Detection: A Multistage Emotion Self-Report Labelling Protocol	135
6.1	Introduction	136
6.1.1	Approach	137
6.1.2	Research Questions	138
6.2	Related Work	139
6.2.1	Emotion Self-Report	139
6.2.2	Characteristics of Emotion Dynamics	141
6.2.3	Labelling and Timing	141
6.2.4	Emotion Elicitation	141
6.3	Data Collection Protocol	142
6.3.1	Participant Task 1: Primary Emotion Activity (PEA) . .	142
6.3.2	Participant Task 2: Emotion Word Calibration (EWC) .	143
6.3.3	Participant and Researcher Task 3: Calibrated Interview → Timeline with Calibrated Words	143

6.3.4	Participant Task 4: Continuous Annotation (CA)	144
6.3.5	Task Order	144
6.4	Exploring Multi-Pass Emotion Self-Reports	144
6.4.1	Commonality in Interpreting Emotion Words	144
6.4.2	Self-Report Modality Consistency via Time Series	145
6.4.3	Comparing Motion Characteristics of Emotion Dynamics	149
6.5	Discussion	150
6.5.1	Multi-Pass and Personalized Emotion Reporting	150
6.5.2	Incorporating Dynamics into Emotion Models	151
6.5.3	Protocol Reflections	152
6.5.4	Future Directions	152
6.6	Conclusion	154
7	Dynamic Emotion Modelling on Incidental Emotion via Videogame	
	Play Controls	155
7.1	Introduction	156
7.2	Background	158
7.3	Dataset Description	160
7.4	Methods	163
7.4.1	Data Instances: Labels and Window Lengths	164
7.4.2	Force Sensitive Resistor (FSR) Data:	166
7.4.3	EEG Data	167
7.4.4	Classification Model Implementation	167
7.5	Classification Performance by Modality	169
7.6	FSR Feature Analysis	170
7.7	Discussion and Future Work	171
7.7.1	Real-Time Predictors of Dynamic Emotion	171
7.7.2	Building Effective Models for Dynamic Emotion Prediction	172
7.8	Conclusion	173
8	Collecting and Labelling Training Data for Dynamic Emotion Clas-	
	sification: A Proof-of-Concept	175
8.1	Introduction	176

8.2	Background	178
8.2.1	Ecological Validity of Emotion in HRI	179
8.2.2	Protocol for Eliciting and Labelling Dynamic Emotion . .	179
8.2.3	Spontaneous Emotion in Training Data for Machine Clas- sification	180
8.2.4	Comparing Modalities	181
8.3	Naturalistic Data Collection Procedure	181
8.3.1	Software and Interface	181
8.3.2	Recruitment Summary	184
8.3.3	Ongoing Consent Practice	184
8.3.4	Setup	185
8.3.5	Data Collection Protocol	186
8.3.6	Dataset Description	189
8.4	Participant Experience	190
8.4.1	Consent Process	190
8.4.2	Observations	191
8.4.3	Questionnaire Responses	192
8.5	Model Training Summary	193
8.5.1	Datastream Pre-processing	193
8.5.2	Feature Extraction and Selection	194
8.5.3	Label Extraction	194
8.5.4	Model Exploration Summary	195
8.6	Limitations and Future Work	197
8.6.1	More Participant Data	197
8.6.2	Emotion Specificity	198
8.6.3	Prompting Emotion Evolution	199
8.6.4	Data Collection and Training in Use	200
8.7	Conclusion	200
9	Conclusions and Reflections	202
9.1	Accounting for the Neurophysiology of Touch	203
9.2	Designing for Emotion Reactivity	203
9.3	Designing for Accountability and Trust	204

9.4	Designing for Engagement	205
9.5	Designing for Specific Care Contexts	206
9.6	Summary	208
Bibliography		211

List of Tables

Table 1.1	A sample set of devices toward interactive machines (* indicates interactivity intended for future work)	10
Table 1.1	A sample set of devices toward interactive machines (* indicates interactivity intended for future work)	11
Table 2.1	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 <i>Problem 1</i> from <i>Related Work</i>	30
Table 2.2	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.	37
Table 3.1	Editable parameters for haptic message design	51

Table 3.2	Pilot results: Test Stimulus Perception. Values are the number of pilot participants able to draw the exact (no partial credit) number of Segments, Direction of motion, and overall track Shape of the test sensation as delivered by the haptic device (Dev) or control human researcher (Hu). Each participant group contained $N = 6$ individuals for 48 trials (8 test stimuli x 6 people).	53
Table 3.3	Eight categories of crowdsourced messages to send to loved ones from a survey of people ($N = 201$) in long-distance relationships	55
Table 3.4	Scenario Prompts for Haptic Message Design and the Number (#) of Designs for Each. Participants designed at most 1 encoding per scenario (some ran out of time before completing all 10 prompts).	56
Table 3.5	Summary of Features Extracted. Parameters expressed in bold-type are vectors – scalars otherwise.	60
Table 3.6	Three raters determined that all 17 wildcard message design strategies fell in three categories, with illustrative examples. . .	64
Table 3.7	Wildcard Messages designed for and interpreted by partners. Participants designed one wildcard message each.	68
Table 5.1	Experimental procedure and data acquisition.	109
Table 5.2	Summary of features extracted from <i>touch</i> , <i>gaze</i> , and select <i>biometric</i> signals.	111
Table 5.3	Data instance count by Emotion and Participant.	113
Table 5.4	A motivating overview of analysis factors.	115
Table 5.5	Weighted F1-scores from 20-fold cross validation varying factors of Gap(+/-), Participant Labels(+/-), and Window Sizes (0.2s, 0.5s, 1s, 2s) on touch T , gaze G , and biometric B features, classifying emotion ($25\% \leq \text{chance} < 50\%$). Classification accuracy is within 0.003 from these values. Weighted F1-scores that are from 0.01 to 0.03 below classification accuracy are indicated with *.	117

Table 5.6	Overall classification performance across all test conditions and modality combinations by accuracy and weighted F1-scores. .	118
Table 7.1	Full list of Calibrated Words used by at least one Participant in their TwCW.	165
Table 7.2	Hierarchical classification scores for each (W)indow / (M)odality where the best combination is 5s-FSR . All W/M models exceed chance by ~2-4x.	169
Table 8.1	Storytelling Prompts and the Most Prominent Emotions Elicited as Reported by Participants. Mean(SD) duration of all stories is 9:30 (4:18).	187
Table 8.2	Relived Emotion Ratings. Participants with multiple sessions are denoted P#-S where S is the session number (P3-1 indicates participant 3's first storytelling session).	193

List of Figures

Figure 1.1	Agents allow for bidirectional influence wherein both the machine and Alice affect one another's expressions.	4
Figure 1.2	A machine that acts as a conduit receives emotional touch from one party, Alice, and conveys it to another, Bob. The recipient, Bob, interprets Alice's intended message based on the machine conveyed experience.	6
Figure 1.3	Machines that influence present an expression on Bob which influences and/or changes Bob's experience.	7
Figure 1.4	Witness machines receives expressions from Alice and tracks, interprets and/or displays Alice's experience.	8
Figure 2.1	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).	24

Figure 3.1	Our tactile animation prototype and participant-designed messages. A touchscreen interface (a) allows senders to draw a track (b) modulated over an 8-tactor array (shown flipped on contact side). Recipients could (c) experience the haptic design, interpolated smoothly between tactors as drawn. In our study, participants designed messages for a close partner: for example, (d) P07b sent a haptic pictogram – though P07a didn’t speak of the sensation in visual terms as puzzle pieces, they did interpret it as <i>connection</i> based on the retracing of a similar path (at the join). (e) P06b created an abstract, rhythm-based sensation from which partner P06a inferred as <i>irritation</i>	41
Figure 3.2	An iterative process (dashed lines) of developing a device suitable for a haptic messaging application. For this paper (solid arrows), we built and piloted a wearable device and conducted a 3-phase haptic messaging study based on designing and interpreting haptic sensations rooted in emotion-laden scenarios commonly experienced by members of close long-distance relationships.	48
Figure 3.3	Representative perceptibility pilot results. A comparison of the test design (L) and a participant-drawn interpretation (R) from each of the three images above (chosen from the 8 message trajectories as depicted in Table 3.2). Each continuous segment is labelled with the order of its playback.	51
Figure 3.4	Participants designed messages of unspecified duration where calm has the largest variation in duration and anger the shortest.	61

Figure 3.5	Designer-created Message Parameters of Track Length, Brush Diameter, and Vibration Frequency for each Emotion Prompt (cross-reference by emotion word for prompt in Table 3.4), including the wild-card message which participants created for their partners. Here, we see that <i>calm</i> tends to small and slow designs with small brush size and track lengths and low vibration frequency; in contrast, <i>attention</i> has a large range of vibration frequencies and brush sizes though mostly small to medium track lengths.	62
Figure 3.6	Confusion Matrices Comparing Interpretation Accuracy of Affective Content for each Haptic Message (count of interpretation instances). In order of increasing accuracy, by (a) human strangers, (b) machine stranger (Random forest classification), (c) designer’s partner, and (d) the designer themselves, a week later; chance = 10%. Red values indicate where the highest mis-classification rate matches or exceeds the diagonal.	66
Figure 3.7	Interpretation Accuracy (%) by Message and Relationship, ordered by decreasing overall recognition accuracy.	67
Figure 3.8	Two very distinct ways of designing for the same <i>excited</i> message prompt.	70
Figure 3.9	P02b particularly enjoyed designing haptic messages after a first try on <i>anger</i> , and imagines developing a vocabulary. . . .	74
Figure 4.1	The robot structure (top right) and a diagram of the participant experience of watching a fear validated video while holding a fur-covered robot that demonstrated one of three breathing patterns (bottom right) manipulated between participants. . . .	83

Figure 4.2	Visualization of the order of videos presented during the procedure and corresponding robot breathing via the manipulation of simple symmetric sine waves throughout the 4 minute and 48 second procedure. The apparent breathing rate in the fear condition plateaued between 220 seconds and 245 seconds, due to mechanical limitations of the robot motor prohibiting it from moving at a faster pace.	85
Figure 4.3	Physical setup showing room layout and relative positioning for participant and researcher over all stages of data collection.	90
Figure 4.4	Locally Estimated Scatterplot Smoothing (LOESS) lines outlining changes in HR over time (solid line), and manipulated breathing pace of the robot (dashed line) over time, in the Fearful Breathing (top), No Breathing (middle), and Stable Breathing (bottom) conditions. Ribbons indicate 95% Confidence Intervals around local estimates. Note: These data are a combination of data presented in Figures 4.2 and 4.3. The Y-axis on the left corresponds to the participant's HR, whereas the Y-axis on the right corresponds to the robot's breathing rate.	91
Figure 5.1	Study setup overview: robot description and participant experience. (a) The robot was constructed from pliant plastic sheets actuated by a pulley, covered with a custom touch sensor, then jacketed in furry fabric to invite touch [37]. It was stationary during the study to eliminate reaction to robot motion. (b) A participant sits supported by pillows and facing the gaze tracker, one hand on the sensor-clad, stationary robot, biometric sensors on chest (RR), thumb (BVP), and index / ring fingers (SC) of resting hand. (c) A schematic of the study room, depicting camera locations relative to where the participant sits by the robot platform.	107

Figure 5.2	Emotion classification accuracy rates from 20 fold cross-validation by <i>modality</i> (Touch + Gaze, Touch only, and Gaze only), <i>window size</i> (0.2s, 0.5s, 1s, 2s), as weighted averages from Table 5.5. Comparisons are also made between having participant labels included (b) & (d) vs excluded (a) & (c), and where 2s gaps are imposed to simulate data loss (a) & (b) vs no gaps (c) & (d). Including biometric data consistently achieves 90-100% accuracy across windows, labels, and gaps (accuracy dips only under the sparsest data conditions: gapped-2s window cases, regardless of whether subject labels are present).	120
Figure 5.3	Comparing how each classification task performed by emotion using touch and gaze features. For subject independent analysis (c) we trained 2 RFs—trained on <i>Excited-Depressed</i> and <i>Stressed-Relaxed</i> separately (no between-set classification – blank entry for <i>Depressed-Stressed</i>). In contrast, a single RF was trained on all 4 emotions in (d).	121
Figure 5.4	Feature selection count by statistic as ranked by Weka’s Best First Attribute Evaluator. Selection % represents how often the feature is selected for use in 100 iterations of 20-fold CV. The dark box for Touch Distribution-Location x Median indicates that this feature is selected 100% of the time; white boxes indicate features that were never selected.	122
Figure 5.5	Changes in individual’s self-report of emotion after Neutralization (start) and Emotion tasks (finish); N=14 for <i>Stressed & Relaxed</i> and N=16 for <i>Depressed & Excited</i> . Overall, we see a move from the origin to the representative quadrant. <i>Stressed</i> and <i>Excited</i> show the strongest overall change along both Arousal and Valence axes. <i>Relaxed</i> shows the least change with disconnected points referring to “no change” from neutral state.	124
Figure 6.1	Roadmap for developing an emotion-prediction engine for an emotionally responsive application.	137

Figure 6.2	Participant tasks and resulting data. At lower left is an EWC example: word stickers placed on a Relaxed-Stressed scale, plus P09's other annotations. The latter resulted from P09 later contextualizing their in-game experience.	140
Figure 6.3	Rating variance by calibration word, ordered by number of participants who provided a rating for that word.	146
Figure 6.4	Comparison of summary statistics and histograms by emotion parameter.	147
Figure 7.1	An emotion experience trajectory estimated by emotion transition. We built models on two modalities: brain activity (EEG) and keypress force (FSR), distinguishing intensifying(+), stable(0), or resolving(-) stress, at 0.5s and 5s windows.	157
Figure 7.2	Our hierarchical machine learning framework employs a two-tiered classification strategy. Initially, a local multi-class classifier is deployed at the parent node level to identify the primary category, termed as the "Calibrated Word." Subsequently, for each emotion word identified in the first step, dedicated models are trained. These models are designed to predict one of three potential outcomes related to the "Stressed" category. This architecture allows for a nuanced understanding of the data by first broadly categorizing the input and then applying specialized models for a detailed analysis within each category.	164
Figure 7.3	Pipeline for model selection and evaluation. We performed grid search CV ($k = 5$) on the training set to tune hyperparameters and select best-fit models for FSR data. The models were then evaluated on an unseen test set to calculate performance metrics. We repeated this process 30 times per participant, and report mean test scores across the 30 runs and 16 participants.	166

Figure 7.4	Structure of the EEG CNN model for classification where each convolution layer uses a 3×3 kernel (of depth 32 and 16 respectively) followed by a ReLU activation function. The inputs to the model are the $5 \times 64 \times 64$ AsMap features [1], while the output is the class output ($N = 3$).	168
Figure 7.5	Relative feature performance by window size. Darker cells indicate frequent selection of better-performing features. The RIGHT directional key is used to advance the character – and game storyline – through the side-scrolling game. A5 corresponds to the sum of the pressure across all keys, while A6 corresponds to the max force over all keys.	170
Figure 8.1	Interface screenshots showing (a) live sensor visualization; (b) emotion word calibration; (c) interview annotation; and (d) continuous annotation stages.	183
Figure 8.2	Physical setup showing room layout and relative positioning for participant and researcher over all stages of data collection.	185
Figure 8.3	Pipeline for model selection and evaluation. We performed grid search CV ($k = 5$) on the training set to tune hyperparameters and select best-fit models for the touch data. The models were then evaluated on an unseen test set to calculate performance metrics. We repeated this process 8 times per participant-session, and report mean test scores across the 8 runs and 10 sessions.	196

List of Abbreviations

AI	Artificial Intelligence
DEAP	Dataset for Emotion Analysis using Physiological Signals
EDA	Electrodermal Activity
EEG	Electroencephalography
EWMA	Exponential Weighted Moving Average
FSR	Force Sensitive Resistor
GSR	Galvanic Skin Response
HR	Heart Rate
HRI	Human-Robot Interaction
OT	Occupational Therapist
TWCW	Timeline with Calibrated Words

Acknowledgments

The existence of this work is due to the support and effort of so many people.

For the expert supervision by Dr. Karon E. MacLean, and committee members Dr. Cristina Conati and Dr. Beck Todd, thank you for your patient guidance.

It has been my privilege to have engaged with numerous collaborators, mentors and SPIN teammates past and present, learning from and with them: Rubia Guerra, Bereket Guta, Paul Bucci, Preeti Vyas, Devyani McLaren, Tommy Nguyen, Rocco Ruan, Hannah Elbaggari, Hanieh Shakeri, Unma Desai, Kattie Sepehri, Dr. Zak Witkower, Dr. Jussi Rantala, Dr. Merel Jung, Yuna Watanabe, Laura Rodgers, Hailey Mah, Qianqian Feng, Anushka Agrawal, Shinmin Hsu, Chuxuan Zhang, Haley Foladare, Jessica Wong, Dr. Yanan Sun, Dr. Soheil Kianzad, Dr. Hasti Seifi, Dr. Oliver Schneider, Dr. Ali Israr, and Dr. James Kryklywy. Thank you for your friendship, encouragement, and late night figure wrangling.

To my family, Hui Wang and Alice Cang and Sandra Trujillo, who have provided endless investment and support: thank you for your gifts of time.

Oliver, Lucy, and Elena: you have enriched my life, making everything else so fun that I had to wait until you were asleep to do most of this writing <3

I am ever grateful to Willow Pavilion and Canuck Place for opening up their doors to me. And to Dr. Lillian Hung, Dr. Hal Siden, Kay, Mario, and Jonah for teaching me how simple it can be to make meaningful connections – your openness and compassion has shown the multitude of forms that caring can take. I am a better parent, daughter, partner, teacher, student, friend, colleague, human being for it.

And to my eight-year old neighbour who had encouraging words for me even while he fights his own big fight: I'm in your corner too, Eric.

Thank you for being my community; how did I get so lucky?

By chapter, my co-authors and I make the following acknowledgments, as published in their respective papers:

- Chapter 3: This project was funded by Meta Reality Labs. We are grateful to the many people who have supported the production of this work: Taylor Bundy and Casey Brown with data collection; Blaise Ritchie with the prototype's design interface; Rubia Guerra and Hannah Elbaggari with figures and reporting analysis; and many others for their supportive input, constructive comments, and helpful edits!
- Chapter 5: We thank Dr. Jessica Tracy for directing us to relived memories as an emotion elicitation strategy, and Merel Jung for her valuable input in developing the methodology. This work was funded in part by Natural Sciences and Engineering Research Council of Canada (NSERC) and the Academy of Finland project Haptic Gaze Interaction (decision # 260026). The study was conducted under UBC Ethics #H15-02611.
- Chapter 6: We thank Dr Rebecca Todd and Dr James Kryklywy for the valuable insight into the neuropsychological effects of emotion evaluation that informed the design of this protocol. Many people have invested time and effort into this project: Kevin Chow, Tyler Malloy, Devyani McLaren, Andrew Moore, Drishti Rawat, Zefan Sramek, Sherry Yuan, and Hafsa Zahid. This work has benefited significantly from their involvement. This work was funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and conducted under UBC Ethics #H15-02611.
- Chapter 7: Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding this work. Human user research was conducted under UBC Ethics #H15-02611.

Dedication

To Lucy and Elena,
who have spent their lives so far touching everything in sight:
you make my world brighter and my days longer.
Thank you for making my story richer.

Chapter 1

Introduction

The pursuit of quality relationships is a fundamental human need [22], so it is no wonder that forming emotional connections make up a large part of our lives [184]. This community socialization involves many nonverbal cues (such as facial expressions, body posture, touch interaction, and eye gaze behaviour), communicating complex and significant information about our present emotion conditions as well as future actions, and can serve as the cornerstone of our personal and professional relationships [29, 105].

We are particularly interested in touch as a driver of closeness [16, 197] and as a proven avenue for emotion communication under both direct contact [115] and machine-mediation [11, 131, 294]. For the most vulnerable among us, touch is a powerful channel for communicating feelings of care, honour, and respect, particularly with our loved ones requiring complex care. We can imagine scenarios where technologically-mediated touch may be able to convey information about emotional needs and experiences, giving caregivers more context clues on how to enhance socio-emotional engagement and possibly improve overall quality of life. However, before we can entrust machines with touch-centric mediation of emotion, we must trust that machines can recognize enough about our highly time-varying and extremely individualistic expressions of emotion to be believably helpful in computer-supported interactive situations.

This thesis explores the present day and beyond of machine assistance for touch-centred interpersonal relationship-building and emotion self-regulation. At

the highest level, we ask: *How can we enable machines to recognize true and spontaneous evolving emotion expressed through touch?* We approach answering this question from two perspectives: first, by articulating and considering a range of roles machine agents must take on in order to engage with human expressions of emotion; and second, planning for future interactivity by breaking down the necessary steps for machine recognition of true and evolving human emotion.

In Part I of this thesis, we begin by examining how emotions are contextualized for in-lab data collection and reporting. With the challenges of describing emotions in mind, we then consider the narrative position of machines in emotion-embedded interaction (1) between people as a communication **conduit**, (2) as an agent with **influence** on an emotion experience, or (3) as a silent **witness** of emotion expression. Emotionally receptive devices have the potential to be emotionally **interactive** where users have emotional responses to a machine's actions or expect their own expressions to have an impact on a machine's 'emotions' [148].

In Part II, we consider that machines designed to mediate and positively reinforce relationships with ourselves and others via touch interaction require the ability to recognize emotionally informative features in users' touch gestures, and the evolution of these features over time, along with the emotion itself. Machine recognition of emotion and/or machine-conduits of emotion that center touch need a system for building personalized models reflective of users' expression and responses. Sometimes, users may have behaviour and preferences that they themselves may not yet be aware of. One challenge is in developing an intuitive and minimally imposing procedure for the collecting and labelling of genuine and spontaneously occurring emotion expression. Such embedded emotion classification models should eventually be capable of spanning the wide range of the user's evolving emotion experience. To better understand how to leverage machine-mediated emotional touch interaction, we examine the workflow of machine recognition of emotion as communicated through touch gesture, addressing the challenges in eliciting authentic and spontaneous emotionally-expressive touch in the lab, accurately labelling these gestures, and finally building computational individualized models.

1.1 A Note on Terminology

There are many ways that emotion-laden touch is discussed throughout this thesis. To ensure consistent understanding throughout, we make clear our definitions behind key terminology used throughout.

1.1.1 Social, Affiliative and Affective Touch

In the context of human-computer interaction and affective computing, we use the term “social touch” to refer to the interpersonal, affiliative, and emotionally-expressive forms of tactile interaction. This is distinct from more functional or utilitarian forms of touch, such as those used for object manipulation or task-oriented interactions. Affiliative touch encompasses the gentle, caressing, and emotionally-laden forms of physical contact that play a crucial role in pre-cognitive social bonding, emotional communication, and the expression of care and intimacy between individuals [167]. We may also distinguish affiliative touch from more general ‘affective touch’ so called to refer to all emotion-embedded touch which also includes negatively-valenced physical contact that may not qualify as gentle. These specialized forms of touch are the focus of our investigation into the design of authentic touch experiences in interactive systems.

1.1.2 Authentic and Spontaneous Emotion

While both spontaneous – naturally occurring, pre-cognitively determined [93] – and posed – acted, expected, requires cognitive decision [148] – expressions of touch may convey emotion, they may well differ in actual performance. Spontaneous behaviour, though more nuanced and difficult to decipher, are seen as more authentic to a person’s true touch performance as they arise naturally from the emotional experience. Posed expressions, on the other hand, may not fully reflect inner emotions due to a disconnect between outward display and internal state [148]. We refer to ‘true’ or ‘authentic’ and ‘spontaneous’ emotion as the natural, unprompted feelings that arise ‘in-the-wild’ [72].

1.2 Reading Emotional Touch

We ground our consideration of machine-mediated physical communication of emotion on three premises:

1. Humans are emotional creatures who colour most interactions with some amount of feeling, whether with other humans [17] or machines [148]. The design of devices and interactions should account for instinctual human emotion [206].
2. People may default to interacting with machine agents the way that they do with other living agents [88].
3. Touch supports and reinforces important pro-social and emotional individual and interpersonal development [16, 113].

Based on these premises, we examine how machines can be used to facilitate emotional interactions in a variety of roles. We see this work as contributing to the future development of a fully interactive machine agent.

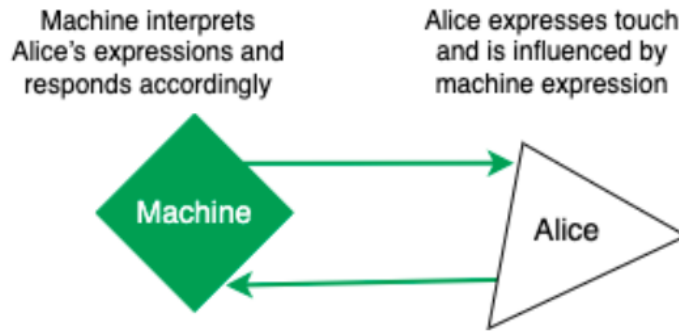


Figure 1.1: Agents allow for bidirectional influence wherein both the machine and Alice affect one another's expressions.

Autonomous machine agents capable of touch interaction include some intriguing devices offering valuable insight. Robots like MIT's Huggable [281] and the commercially available Paro [306] have been used to leverage the comfort afforded by stuffed animals but also making them interactive to study how these

devices could be used for therapeutic and clinical care. While these systems have some limited responsiveness by dint of providing predefined responses to particular stimuli, they are not able to respond to an estimate of a user's emotional touch expression, as a real being might, as they lack the ability to interpret this expression. The Huggable has a teddy bear form factor primarily designed for young children experiencing stress and pain of hospitalization. With an accompanying smartphone application, the Huggable can also be an extension of a human tele-operator with the phone's speaker and screen being co-opted to play the operator's voice and display the bear's eye movements and expressions [141], making it appear autonomous. By contrast, the Paro is a standalone therapy robot in a baby harp seal form factor with limited motion capabilities focused at the large blinking eyelids, neck, and flippers. Intended for elder care, the Paro is cute and has babyish features, inviting attention in the same way animals might to facilitate more social engagement [132, 262].

The challenge of 'closing the loop' for autonomous emotion communication still exists for machines that receive, and communicate with, natural human emotion expression. To better understand the processing pipeline required to complete the interactive loop in realtime, devices have been developed to examine emotion interaction in distinct stages: (1) to 'witness' or sense user emotion and then (2) to generate an response with human-interpretable emotional content. The Haptic Creature [319], as an example, was developed with the plan that users' touch behaviour would trigger a robot response which was itself designed to help calm them [253]. The calming should in turn be detectable through changes in the users' touch, and ongoing sensing of it [321]. The authors identified each of these stages (sensing, recognition, emotion rendering) as an explicit research challenge – and indeed, the community continues to work on all of them more than 15 years later. The professionally engineered CuddleBot [4] was designed to be computationally capable of handling the recognition task as well as displaying physical expression suitable for an ongoing interactive context (as opposed to simply relaying static emotional snapshots). Similarly, the weight-shifting OMOY has been studied for how people interpret this modality of emotion expression with future work describing the interactive content where OMOY's behaviour could influence users' emotions [204].

In the following, we examine both stages of the interactive loop separately and then in concert (how humans communicate emotion through machine-mediation). In doing so, we describe three emotionally purposeful roles that machines can take on and provide examples of each.

1.2.1 As Conduit

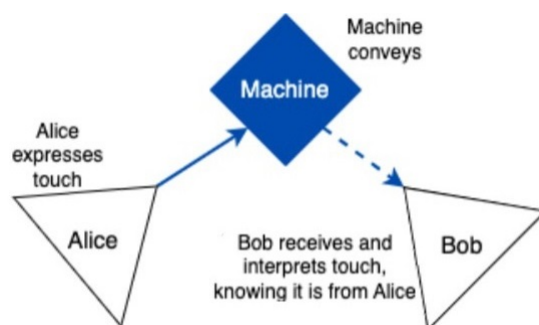


Figure 1.2: A machine that acts as a conduit receives emotional touch from one party, Alice, and conveys it to another, Bob. The recipient, Bob, interprets Alice’s intended message based on the machine conveyed experience.

The global pandemic gave society unprecedented and widespread experience with physical distance and restricted direct contact with loved ones. Even without quarantine measures, separation between family and friends for extensive time periods has been on the rise [71, 242] as our lives can take significant turns for professional, education, military, or health reasons, sometimes requiring extensive time away from others. While audio- and video-calling can be used to stay connected, touch technology today offers fewer opportunities for bridging physical distance. Relatively recently, machines able to act as emotional conduits – sometimes dubbed *affective technotouch* for emotional embodied encounters [68] – have been developed for commercial and research purposes. Receiving cues from one person and transmitting them to another, these devices are designed to communicate emotion. They can take on a variety of form factors, from commercial wearables in the form of bracelets like the [Hey](https://feelhey.com/)¹ and [Bond Touch](#)² that transmit squeezes and vi-

¹Hey product descriptions at <https://feelhey.com/>

bration respectively between partners, to research devices featuring sensations like vibration motors [228] and haptic knobs [275] to communicate a range of emotions. More specific use cases of emotion communication include stress transmission using force feedback devices [93], or thermal manipulation as in the Nakama teddy bear [311] to send feelings of warmth from parents to their children; even mid-air contactless sensations can contain interpretable emotion content [207].

Considering the real-world scenario of machine-mediated communication in other forms, we posit that haptic message exchanges are also likely to occur under specific contexts. The messages people design are likely to reflect the emotion they are currently actively experiencing. For instance, Betty just heard great news and wants to alert her partner that they should celebrate or Jon noticed the dishes are still not done despite reminding their partner twice the night before. In the study described in Chapter 3, we brought in couples to create ‘haptic messages’ with specified emotional intent to send to their partner. By comparing how message interpretation performance varies by relationship closeness (strangers, partners, selves one week later), we can better examine the design and interpretation of machine-mediated emotional touch as represented by vibrotactile signals.

1.2.2 As Influence

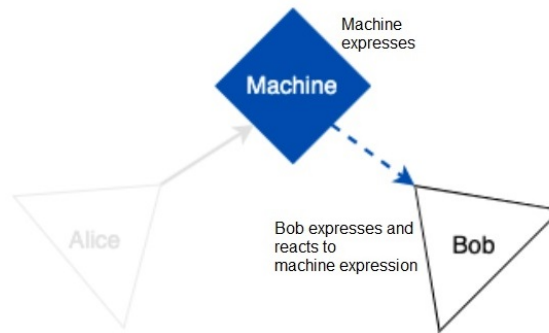


Figure 1.3: Machines that influence present an expression on Bob which influences and/or changes Bob’s experience.

From here, we zoom in to focus on human-machine interaction starting with

²Bond Touch product descriptions at <https://uk.bond-touch.com/>

examining devices designed to influence a human user’s emotional experience. In the case of the Haptic Creature [319], the lap pet-sized robot demonstrated that displaying slow calm breathing behaviours via movable plates on its Force Sensitive Resistor (FSR)-embedded fibreglass body can affect stress-related physiological measures. People holding the robot and experiencing 20 breaths per minute against their chest felt a reduction in their heart and respiratory rates as well as self-reported anxiety, suggesting altogether a positive shift in emotion valence [253]. Slow and rhythmic haptic displays promoting calm and relaxation is corroborated by the Heartbeat Cube [327], also held in the arms by users, as well as the Affective Sleeve [215], a network of shape memory alloys worn on the forearm.

We wondered how effective simple robot motion can be in influencing the experience of humans in close contact. To answer this, we devised a study with three breathing patterns that participants would feel while holding a robot close to their chest and compared the human physiological response during a fear-eliciting stimulus (Chapter 4).

1.2.3 As Witness

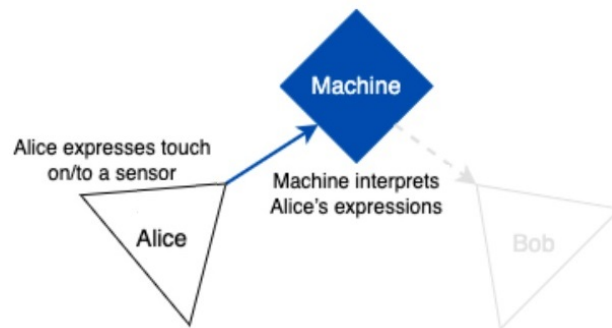


Figure 1.4: Witness machines receives expressions from Alice and tracks, interprets and/or displays Alice’s experience.

Devices that witness our behaviour (i.e., observe or track it) are increasingly common with fitness and physiological trackers becoming ubiquitous in wearables like smart watches and Fitbits, even step tracking applications on phones to help promote healthy activity and performance. Similar devices can also be provided

with the ability to act as a witness for our emotional lives. Some are designed to detect, and directly convey, incidental biofeedback like the Breeze necklace where sensors in the pendant are in contact with users' chests; others infer emotional content from subconsciously generated behaviour, as in the Haptic Creature [5], whose FSRs read user touch and use machine learning to convert these pressure readings to emotion prediction.

Devices can also derive emotional information from utilitarian movements that we don't always think of as emotionally expressive like that of typing on a keyboard or navigating with a computer mouse. Simple computer keyboards can be equipped with pressure sensors on the keys and these pressure readings reveal the typist's stress at 93.4% accuracy [188]. In fact, device usage can be used to predict emotion even without modification. Regular point-and-click or drag-and-drop interactions with a computer mouse can produce metrics that are well-correlated to biophysical markers of stress [286]. Typing behaviour on EmoKey, a keyboard app for an Android smartphone with a capacitive touchscreen, can differentiate between four emotions (happy, sad, stressed, relaxed) with 78% accuracy [98].

Since emotion expressions may differ in a lab collection *vs.* in a real-world context, for a device to recognize real-world emotions, it must be trained with the same spontaneously occurring emotion data. We use an emotion elicitation technique called 'emotion recall' or 'relived emotion' where participants use stories from their own lives to bring up true emotion [57], building and evaluating recognition performance from this emotion data.

Developing machines to respond to human emotion requires that they are built to be sensitive to human emotion cues in real-life contexts. Throughout this thesis, we advance the capability of an emotion recognition engine by exploring how to collect and label naturally and spontaneously evolving emotion.

Table 1.1: A sample set of devices toward interactive machines (* indicates interactivity intended for future work)

Role	System	Form	User Experience	Haptic Sensing	Interpretation of Haptic Input
INTERACTIVE AGENT	Paro [132, 268]	~6lb-furry seal robot	Globally commercialized for older people with dementia; ↓ agitation, anxiety, depression; ↑ social engagement, mood, quality of care experience;	Microphone, whisker tactile sensors, touch sensor on head, jaw, back, flank, flippers, position/orientation	Force and locale of touch maps to coordinated response on eyelids, neck, flippers, vocalizations
	Huggable [141]	Plush teddy bear	Tele-operated using smartphone app; promotes conversations, positive interactions (vs. teddy bear, virtual agent) with children	Pressure + capacitive sensors	Wizard-of-Oz responses to haptic input (non-autonomous)
	OMOY* weight shifting robot [204]	Handheld humanoid robot embedded with movable weight	Hold OMOY to experience weight-shift patterns (over 4 parameters), weight moving quickly to the front	None used at present	Intended for future interactivity, present work focuses on user perception of robot expression
	CuddleBot	Furry lap pet	Social touch gestures trigger coordinated behaviours from motors in neck, ribs, back	Fabric touch sensor	7 social touch gestures at 90~95% recognition when robot was not moving (degrades to 79~86% when in motion)

Table 1.1: A sample set of devices toward interactive machines (* indicates interactivity intended for future work)

Role	System	Form	User Experience	Haptic Sensing	Interpretation of Haptic Input
CONDUIT	Bond Touch [68]	Bracelet	Paired bracelets; Direct touch transmits to receiver unit; app for customization	Capacitive touch + app	User interpretation of vibration + lights
	Hey Bracelet [68]	Bracelet	Touch triggers a squeeze on paired bracelet via Bluetooth (smartphone app)	Capacitive touch + app	User interpretation of squeezes
	Handheld remote [228]	Remote control	FSRs/capacitive touch pad triggers VT! actuators on partner device	4 FSRs & capacitive touch @ 100Hz	Partner interprets localized vibration on arousal-valence
	1-DOF knob [275]	"Twiddler" haptic knob	Spinning knob triggers force feedback on paired knob	4000 count/rev optical encoder	54% recognition rate overall on 4 emotions spanning affect grid
	Geomagic Touch driving stress [93]	Force sensitive multi-DOF end effector	Directing the end effector controls steering in a driving simulation; movements are recorded and played back	Steering + grip force	user interprets stress by higher speed, jerkiness, and force
	Nakama companion [311, 312]	Teddy bear	Heartbeat and temperature of loved ones can be displayed on bear	HR sensor, thermometer	User rates perceived closeness from transmission of heartbeat + warmth
	UltraHaptics system [47, 207]	Ultrasound screen	Ultrasound waves are directed to users' hands hovering midair over the system	Frequency, intensity, duration	User-defined mid-air haptics is interpreted for arousal + valence
INFLUENCE	Haptic Creature* [253, 319]	Furry lap pet	Users hold a breathing robot with biometrics tracked, self-report emotion response	FSRs	Device calms by ↑ valence, ↓ anxiety, heart rate, respiratory rate
	Heartbeat feedback cube	Handheld cube [327]	Holding the cube 'beating' to user's heartbeat increases relaxation	Stethoscope (+ mic) to vibrotransducer	Feeling own heartbeat affects biofeedback loop ↓ heartbeat; ↑ HRV
	HaNS notification bracelet [289]	Notification timing console	Speaker's bracelet vibrates to cue 3m, 1m, 0m to talk end	FSR cues delivered via timing console	Speakers report cues help ↑ awareness; ↓ distraction
	Affective Sleeve [215]	Shape memory alloy cuff sleeve	Sleeve produces rhythmic haptic action (light pressure + warmth)	Respiratory rate, EDA	Breathing rate increases with actuation tempo; slower rhythm promotes calm
WITNESS	Haptic Creature* [5]	Furry lap pet	Strokes a breathing, purring body	56 FSRs; accelerometer	Emotion classified from touch in 9 states ³
	Breeze breathing biofeedback [90]	Pendant necklace	Strokes and squeezes to activate light, sound, vibration	Onboard accelerometer, gyroscope, magnetometer	Emotion classified as valence, arousal, dominance; labelled by SAM self-report
	Computer mouse [286]	Computer mouse	Regular mouse use: point-and-click, drag-and-drop, steering	Kinematics of mouse movement	Motion patterns to classify stress; HRV, ECG, self-report for ground truth

³Emotions across Russell affect grid: distressed, aroused, excited, miserable, neutral, pleased, depressed, sleepy, relaxed

1.3 Prerequisites for Machine Recognition of Emotion in Touch

In order for machines to be able to engage in affect communication and present reasonable emotional responses, they must possess an architecture for recognizing human emotional expression. Machine recognition of emotion requires a sensing system to receive human-expressed input and a simplified internal model to allow for emotion recognition, in some form of classification or prediction [221]. Here, we consider the components used to create the emotion training data that build classification models. We compile authentic emotion data from spontaneous elicitation and record emotion-encoding modalities for use as model inputs, reporting and labelling as true to the experience as possible.

1.3.1 Real Emotion Elicitation

Generating spontaneous emotion data that is representative of real use is crucial for developing a system that has functional machine recognition once deployed “in the wild” [72]. At the research stage, it would be ideal to collect training data in the lab, where we can control conditions. Unfortunately, emotions elicited in a lab [57] between human and an affective agent [148] can be challenging to generate with high ecological validity and worse, can differ dramatically from naturally occurring or spontaneously experienced emotion [5, 18, 93]. Studies building models of emotion behaviours based on “acting as if” or “imagining that” one feels an emotion [5, 321] may not be representative of true and spontaneous emotion experiences [126].

Relived or recalled emotion using stories from real lives may allow for emotion expression to approximate true-to-life experiences. Elicitation methods outlined by [57, 80, 180] demonstrate that, even in a lab environment, strong physical reactions similar to those occurring from the original incident can be produced.

To consider the viability of computational models of touch data collected in lab, we evaluate the performance of machine learning recognition of touch and other biophysical signals co-occurring during recalled emotion elicitation.

1.3.2 Emotion-Embedding Modalities

While language is an important method for conveying emotion, emotion expression occurs across many other nonverbal channels, so much so that classification models of emotion can be, and have been, built out of a variety of emotion-embedding modalities. Examples are many: facial expression recognition by image [158], brain activity as measured via electroencephalography (Electroencephalography (EEG)) [3], biophysiology data as in electrodermal activity or skin conductance [48], Heart Rate (HR) (heartrate), respiratory rate [77], even eye gaze behaviour can reveal valuable affective information [139, 173].

Why Touch?

We examine touch behaviour as an intentional communication channel in social touch where one agent touches another to convey emotion information [115] – strokes to calm or reassure, taps to direct or request attention. While touch can serve as an incidental signal, such as in typing behaviours [81] or grip force in steering [307], it is also known to encode interpretable emotive content [114, 115] with pressure from both normal and shear forces providing valuable affective information [53, 152].

Touch is one of the earliest modes of communication and reassurance in our lives and is necessary for healthy emotional and prosocial development [113]. To create spontaneous and honest emotion interactions, we consider the potential of employing touch sensing in the form of smart fur [87] or pliant piezoresistive fabrics [41], prioritizing unencumbered and spontaneous interactions [42] where possible.

Technical Sensing Modality Selection

The sensing modality selected for computational model building must suit the application purpose and requirements and have sufficient power to discriminate the emotions relevant to the application. Electroencephalography (EEG) measures brain activity occurring from cognition to neurological response and emotion regulation [109, 210, 227]. It has been well-explored for emotion classification with the brain being closest to the ‘root’ of the emotion experience [176]. Biometric data like heart rate variability (as measured by blood-volume pulse or BVP), skin conductance, respiratory rate – physiological responses coincident with strong emo-

tional experiences – have also been used as strong correlates for emotion expression and thus high-performing signals for emotion classification [60, 104, 155, 221]. However, modelling using these biological signals requires intrusive sensing apparatus, time for subject preparation and calibration, and as collected data is notoriously noisy, also requires extensive data filtering, cleaning, and pre-processing [44].

Some modalities allow for a more out-of-the-box experience (with minimal set-up investment). Neither touch nor gaze detection necessarily requires users to don specialized equipment as sensing can be embedded in or near the device itself. Some commercially available gaze trackers like those built by Tobii (www.tobii.com) are designed to be set at the base of a screen or intended field of view. The tracker can detect pupillary response, gaze distance, focal location, saccade and fixation patterns – all notable features in emotion recognition in gaze behaviour [183, 276, 291].

For use on robots or other motion-enabled agents, touch sensors must cover expected touch areas, which can be quite large depending on the intended interaction. For instance, a robot pet may need to have sensors embedded along the dorsal area, from head to tail, to pick up long petting strokes. If the machine has any motion capability, the sensors or sensing surface may need to be physically flexible and malleable. Networks of FSRs (force sensitive resistors) [319], stretchy piezoresistive fabrics covering touch-intended surfaces [41, 152], or proprietary palm-scale pressure sensors positioned strategically near likely touch points like the head and back [267] are some strategies that can accommodate underlying motion.

Where we embed touch sensing capabilities into devices inviting or requiring touch interaction, there is an opportunity to leverage a natural and intuitive modality for emotion communication and take an unobtrusive look into spontaneous and honest emotion expressions. We evaluate how well pressure-sensing in touch behaviour can encode emotion in both naturalistic touch (Chapter 5) and incidental touch (Chapter 7), with and without the support of gaze and biometric markers, or EEG data capture respectively.

We select these markers for being well-researched as emotion-embedding modalities. While facial expression is another prominent area of emotion recognition study [79], there is a growing body of evidence [97, 138] that facial expressions of emotion are complex and are often modulated to present more positive expressions

– like smiling [17, 122] – to hide more negative experiences. Furthermore, collecting facial expression data would require our devices have embedded camera and vision systems which would impact our computational requirements and introduce additional practical and privacy concerns in terms of when and what to record.

1.3.3 Conceptualization and Labelling

In order to estimate an amorphous quantity like emotion we first choose a representation metaphor which defines how we regard the emotion experience, the language we use to describe it, and the parameters with which we attempt to capture it [38, 169, 211]. Altogether, this descriptive framing and parameterization is sometimes referred to as *emotion modelling* [198]. To avoid overloading the term “model” or confusing the use of emotion models with classification models of emotion, we adopt [38]’s terminology of emotion *metaphor* to refer to how we think about emotion representation. Throughout this thesis, unless specified otherwise, we use *model* to refer to the computational (e.g., machine learning) model implementation rather than theoretical models that provide structure around how we can define and reason about emotions (e.g., Russell’s circumplex model of emotion [237] or appraisal model [84]). In preparation for building computational models of emotion, we start by addressing commonly used emotion metaphors: emotion *states*, *dynamics*, and *appraisal*.

Emotions-as-a-State: Classifying emotion as a single *state* has many practical benefits for machine recognition. There are many validated instruments for identification and measurement of arousal, valence, and dominance dimensions [238] that can be used to distinguish between emotions. Reported emotions can be described on a 2D plane as as in the arousal-valence grid [237] or on independent linear scales as in the Self Assessment Manikin or SAM scale [30]. These are beautifully simple measurement scales which employ forced choice and offer simple and straightforward classes for data labelling. In contrast, emotions rarely fit into convenient boxes. However, our emotional lives are complexly dynamic in situation-dependent ways: who we are with, how recently our physical and emotional needs have been met, and why we are in the present moment with all the baggage of our cultural and personal history [20].

Emotion Dynamics: Emotions evolve throughout the course of a single event or experience, as well as longer extents of time [169]: consider the emotional journey followed by your favorite engaging movie scene. Psychologists Kuppens et al [169] propose dynamic emotion metrics to describe changes across an emotional experience, with the most prominent being *emotion inertia* (resistance to variation, quantified as signal autocorrelation); *emotion instability* (mean square of successive differences as the amount of change); and *emotion variability* (within-subject variance respectively to represent the range of change) [130, 278]. Operationalizing concepts rooted in emotion dynamics for computational applications requires labels capturing transitional emotional experiences as they happen.

We propose the use of *emotion direction* as a dynamic emotion metaphor to describe where a present emotional experience is evolving towards.

Labelling Approaches: Labelling emotion for classification purposes is a challenging activity, operationalized by reducing complex emotional ideas down to low dimensional static elements like Russell’s arousal-valence grid [237] or the Self-Assessment Manikin [30, 246]. While simplifying the label helps with discrete classification, we may be throwing out valuable richness that is necessary for approximating real-world emotion expression [39, 148]. Novel labelling procedures try to build in more richness, capturing more participant introspection and aligning that with researcher observations – examples include reporting of both personal reflections of shame experiences [246] as well as episodes of pain [25].

To harness the richness in reporting, we consider how to incorporate easy numeric or discrete-valued dynamic emotion labelling aligned with the more open-ended personal reflection.

1.4 Thesis Contributions and Organization

To answer the question: “*How can we enable machines to recognize true and spontaneous evolving emotion expressed through touch?*”, we first examine the roles that machines play in emotional interaction with human users, then explore emotion recognition engines that evolve with spontaneous user expression.

Our approach to the problem of creating interactive machine agents to promote realtime emotion communication via affective touch is reflected in the structure of

this thesis. In Part I, we scrutinize how models and metaphors can reflect lived experience (Chapter 2) and through this lens, describe design explorations of how an interactive agent may be able to facilitate touch-based emotion communication with human users: as the **conduit** of emotion between two people (Chapter 3); as a factor of **influence** on a single person’s emotion experience (Chapter 4); and as a **witness** to human expression of emotion (Chapter 5). These works contribute design insights about how machines can or should function in these roles, and how to facilitate them, such as:

1. For machines to react appropriately to human expressions of emotion, they may require the underlying recognition engine to reflect the complex and time-varying nature of the emotional experience (Chapter 2).
2. **For Machine Conduits:** Close intimate partners are better than both strangers and machines in classifying emotional intent behind machine-mediated haptic messages. Given that strangers and machines perform similarly in message interpretation, we posit that context clues and shared personal history may be key factors in understanding the emotional intent behind touch-based digital messages (Chapter 3).
3. **For Machine Influence:** Breathing behaviours in pet-sized robots can influence the person feeling those behaviours, both towards [253], and away from, relaxation (Chapter 4).
4. **For Machine Witnesses:** Touch behaviour exhibited during an emotional experience may be sufficiently distinct as to be machine discernible (Chapter 5).

In Part II, we examine one element required of realtime interactivity – dynamic emotion modelling – and propose a data collection methodology (Chapter 6). By using a rich emotion reporting and labelling procedure, we allow for classification of emotion transitions and differentiate whether an experience is *stressful-but-resolving-towards-relaxed* or *stressful-and-getting-worse*. We evaluate this data collection and labelling protocol with incidental affective touch produced from playing a horror video game finding that even incidental touch pressure in key-press force is recognizable as distinct emotion transitions and published the dataset

for community exploration (Chapter 7). Finally, we explore how evolving models of naturalistic emotion communication can be trained for use in touch-sensitive devices leveraging personalized emotion transition recognition. To advance real-time dynamic emotion classification, we introduce a proof-of-concept training pipeline to build models personalized to evolve with the user, from data collection, to rich emotion reporting, to classification modelling parameters (Chapter 8). Contributions here are (a) methodological and (b) the provision of a feasibility assessment:

1. To capture the complex nuance in fast-evolving emotion, multiple labelling stages can be used to provide rich context when tracking and labelling an emotional experience (Chapter 6).
2. As part of Chapter 7, we contribute a novel dataset of brain activity data and incidental touch pressure collected while participants played a horror video game (the FEEL dataset) and inspect classification of dynamic emotion – evolving emotion within a time window (e.g., differentiating between happy-getting-happier *vs.* happy-getting-anxious). Performance from keypress force data (F-1 scores benchmarked at 0.82) encourages our next step: classifying dynamic emotion labels with natural and spontaneous touch in more unconstrained or ecologically valid emotional expression.
3. With a custom tool to help with labelling spontaneous and evolving emotions, we demonstrate a proof-of-concept for how training data can be generated for classification models of real-world ‘in-the-wild’ emotion evolution (Chapter 8).

In our Conclusions and Reflections (Chapter 9), we consider the open design questions that help to advance the responsible development of emotionally responsive devices and machine agents (Chapter 9).

Part I

**Emotion Communication with
Machines**

There is an extensive body of research demonstrating that the emotion communication that we partake in through direct (e.g., person to person) affiliative touch is transferable in machine-mediated touch. Here, we use ‘affiliative touch’ to refer to the gentle, caressing, and emotionally-laden forms of physical contact that play a crucial role in social bonding, emotional communication, and the expression of care and intimacy between individuals [167]. This is distinct from haptic messages, which are a semiotic form of touch communication – transmitting particular meaning [62] – rather than an affiliative one.

To gain a deeper understanding of *how* affective touch conveys meaning between individuals through devices, we start with Chapter 2’s conceptual ideas on metaphorical representation of emotion, the logistical challenges with eliciting and describing authentic human emotion. We then analyze three different devices and their respective use cases. We also explore the roles that the haptic technology plays in these interactions: first, as a **conduit** of emotion-intended touch between human users; second, as an **influence** on one’s emotional experience through haptically salient breathing patterns; and third, as a **witness** to human emotional expression through touch behavior and biophysical signals.

In Chapter 3, we examine an example of a device serving as an emotional **conduit** with machine-mediated emotion-encoded haptic signals transmitted from one user to another. We developed a haptic animation display worn on the forearm to design and display emotion-laden messages created by close partners. We evaluated how well people recognize and interpret these messages under increasing contextual relationship history (i.e., from strangers, their own partners, or themselves a week later) [45].

Chapter 4 describes an investigation into the emotional impact of experiencing breathing expressions and considers the efficacy of the breathing robot platform as a device of emotional **influence**. We examined physiological responses of people as they hold a robot moving in three distinct sinusoidal ‘breathing’ patterns. As first described in the Preface, this work was a collaboration with UBC Psychology and is currently under review.

To round out device interaction, we include a paper as published in 2021 as “Discerning Affect from Touch and Gaze During Interaction with a Robot Pet” (IEEE Transactions on Affective Computing) in Chapter 5. We used a robot form

wrapped in a custom fabric touch sensor that collected touch behaviour as a **witness** of user emotion expression as they shared an emotional moment from their lives. Augmenting the touch data with gaze trackers and biophysiology sensors, we considered how to use these three modalities in emotion recognition (where each story was labelled with a single emotion label).

Here, we examine various devices by the roles they may play alongside human emotion experiences, noting lessons learned as recommendations for device design and classification parameters (e.g., time windows, modality preferences, feature selection). Through the common thread of machine recognition of emotion, we motivate our interest in developing more intuitive and comprehensive models of emotions evolving in realtime as featured in Part II.

Chapter 2

Metaphors for Emotion: An Argument for Rich Emotion Labelling

Summary

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.

2.1 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 [320] 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. 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.

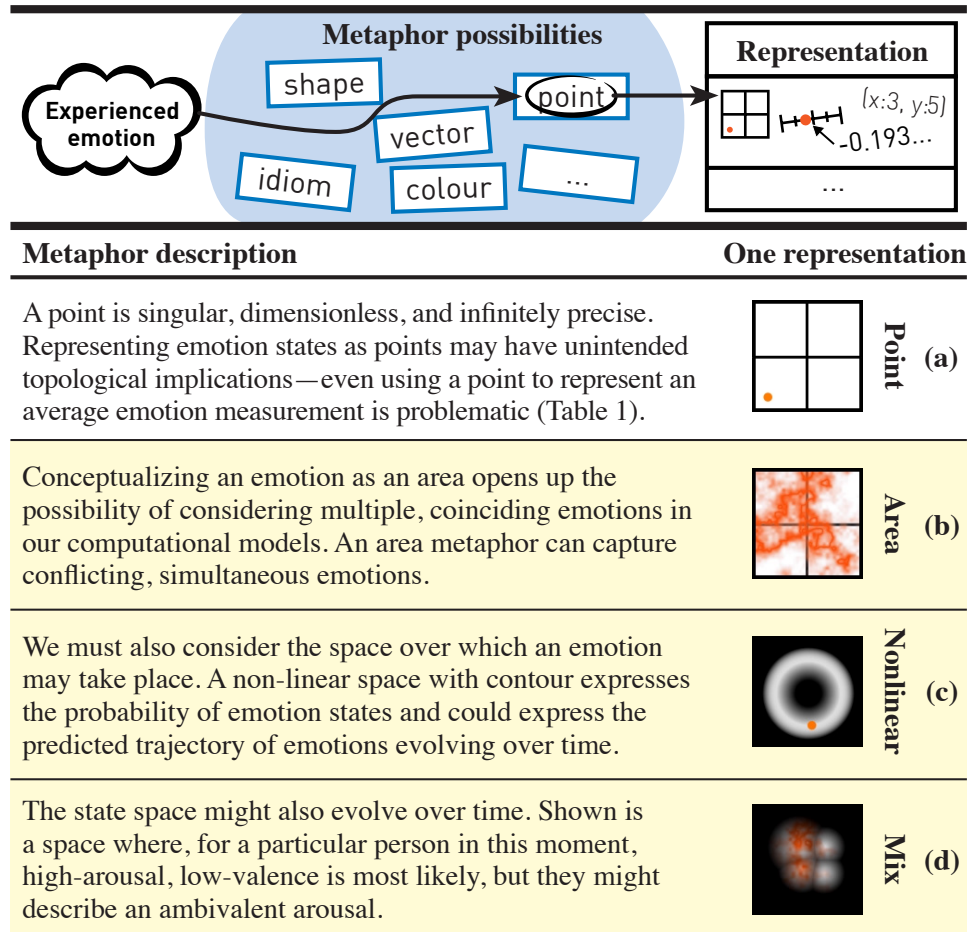


Figure 2.1: Experienced emotions can be reasoned about through the use of metaphors: abstract concepts (mathematical, literary, etc.) that stand in for real-world phenomena. Metaphors can be turned into a multitude of concrete representations to serve different purposes. A common metaphor for emotion 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).

- **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.

2.2 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.

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 [172]. 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 2.1).

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 [172]).

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 2.1, top right) [236]. 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 [19]. For example, to communicate with participants about their emotion, we can employ instruments such as the Affect Grid (a discretized 2D circumplex) [239] or the Self-Assessment Manikin (SAM), which splits the arousal-valence-dominance space into three scales with cartoons for each scale item [30].

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.

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).

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.

2.3 Related Work

Recent theoretical work in emotional interaction has challenged the dominant “signalling paradigm” [148] 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 [148, 175]. 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 [12, 37, 125, 148]. 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 con-

tinuous interaction (or “emotion coordination”) [148]. 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 [86].

Table 2.1: 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*.

Implicit Assumption 1: Emotions can be represented as a single point-like state			
<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.
Implicit Assumption 2: Emotion space is continuous and linear			
<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?

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 [33, 78, 86, 148], Russell's dimensional model of affect [24, 35, 201, 241, 277] or a combination of both [243, 320]. Instruments include the Affect Grid [239], the Self-Assessment Manikin [189, 241], or the PANAS scales [12].

Herein lies the dilemma: computational models of behaviour require labels, but behaviours cannot be consistently and directly labeled with a single emotion [175]. We could add contextual details to computational models to improve labelling accuracy [31, 37, 64]. Alternatively, we could actively choose to represent conflicting or mixed emotions, aligning more closely with known neurobiological phenomena [181] as well as how behaviours are experienced and interpreted in real life [35]. We present a discussion of alternative representations in Section 2.4.

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 [123]. 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 [135]. 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 [265, 283, 314] 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 contrast-

ing questions using different frames [33, 35, 201]. Others establish frame through clarifying interviews where participants explain their interpretation of the study task [37, 175]. Still others use concepts from theatre. Bucci *et al* establish roles, characters, and settings for an interactive scene [37]. Westlund *et al* do this through an interactive theatrical process [310]: 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 [190].

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 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 [285]. 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 [32, 223, 282].

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 or separability. The required introspection to quantify an experience on multidimensional scales – even just 2D as with valence and arousal – may compound dissociation from the lived emotion [164, 165], further obscuring ground truth estimation.

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 [27, 94, 148]. 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 2.6.

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.

2.4 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 2.1) 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' [19]. 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.

2.4.1 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 2.1.

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.

2.4.2 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 2.1 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.

2.4.3 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.

2.5 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 [148]. 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 2.2, 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 [30]. 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 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.

Table 2.2: 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
	Participant (Jan, left) is evaluating how she feels about Robot (Can, right). Jan is being asked to interpret her subjective feelings about how Can is making her feel.
	Jan is evaluating what Can is trying to convey. Jan is being asked to interpret Can's communicative behaviour. Can's expressions give <i>evidence</i> for a hidden subjective state.
	Jan is evaluating how Can feels. Jan is being asked to interpret a set of behaviours over some duration that indicate Can's emotional state.
	Jan is evaluating how Can feels about her. 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.
	Jan is evaluating how she currently feels. Jan is being asked to inspect her body/brain and describe some kind of mixture of mood, emotion, affect, or physiological perceptions.

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.

2.6 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.

2.7 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.

Chapter 3

Machine as Emotion Conduit: An Example of Haptic Messaging in Emotion-Laden Scenarios

Summary

Touch is valued for supporting emotional bonds. How can people access its warmth and nuance remotely, when tech-mediated proxies are so different from direct touch? We assessed the viability of haptic animations as affect-embedded tactile messages, highlighting findings which demonstrate how crucial relationship and shared history is in influencing these expressions in design and interpretation. To investigate haptic messaging, we first identified a set of 10 common emotion-imbued scenarios by surveying 201 people in distance relationships. Then, using a novel prototype of a wearable spatial vibrotactile display, 10 intimate dyads designed 167 haptic encodings matching the provided scenarios plus 17 user-defined “wildcards”. A week later, 21 individuals interpreted sentiment from encodings designed by themselves, a partner or a stranger. We examined design strategies, engagement, and compared human vs. machine interpretation accuracy. A striking finding was participants’ facile use of shared context when it was available, building on “inside stories” to communicate subtle meanings with high effectiveness despite the unfamiliar medium, and doing so with evident fun. We analyze recognition accuracy and share insights on what it might take to make interpersonal

haptic messaging work.



Figure 3.1: Our tactile animation prototype and participant-designed messages. A touchscreen interface

(a) allows senders to draw a track

(b) modulated over an 8-tactor array (shown flipped on contact side).

Recipients could

(c) experience the haptic design, interpolated smoothly between tactors as drawn. In our study, participants designed messages for a close partner: for example,

(d) P07b sent a haptic pictogram – though P07a didn’t speak of the sensation in visual terms as puzzle pieces, they did interpret it as *connection* based on the retracing of a similar path (at the join).

(e) P06b created an abstract, rhythm-based sensation from which partner P06a inferred as *irritation*.

3.1 Introduction

Social touch interactions add nuance to our communication – a light squeeze on an anxious patient’s arm calms them; a firm handshake asserts trust in a newly struck business deal; even a light tap on the shoulder can increase trust and cooperation between strangers [170]. We communicate comfort, love, and safety through touch, promoting pro-social behaviours and forging deep emotional bonds that help form and maintain relationships [16, 197].

It is increasingly common for partners, family members, and close friends to be separated e.g., due to professional, academic, military, health responsibilities [71,

242], fueling a growing appetite for machine-mediated social touch [131, 294] that can re-introduce valuable touch-based interactions where natural person-to-person contact is not practical [142].

How we perceive a communicated sentiment [39, 143] can be heavily influenced by the pre-existing relationship. Natural interactions take place within complex ecosystems of history, condition, and purpose, all of which color the encoding of emotional perception [18]. Studying how these interactions are received and interpreted must include the context and relationship they exist in [21]. This is certainly true of touch: e.g., touch between strangers is unlikely to be interpreted as *surprise, envy, or pride* [114, 296].

People are capable of affectively interpreting simple notification-style tactile sensations [233, 256, 294]; in fact, it is natural to comprehend signals like high frequency choppy buzzing as urgent irritation, or soft rolling rumbles as calming reassurance[255]. Further, haptic animation displays have been embedded into a chair for on-back interaction to incorporate multimodal immersion for visual media [136] and incite a number of intriguing experiences where discrete tapping sensations simulate rain, or low rumblings evoke the purring of a big cat [247]. For this work, we test the feasibility of vibrotactile animation for haptic messaging. To leverage these sensations on a wearable, we ask: can partners with a shared context and history convey high-resolution emotional information, using just low-resolution spatial vibration through a relatively simple vibrotactile animation display?

To test the feasibility of vibrotactile animation for haptic messaging, we prototyped a wearable haptic display, scaling a large chair-sized interface [247] down to an array of 8 small tactors to fit along the forearm (see Figure 3.1). The accompanying message design interface uses the exact principles developed by [248]: users define the sensation by directly drawing on a touchscreen and a continuous tactile signal (i.e., without unintentional segmentation or path break) is interpolated spatiotemporally along the drawn curve. To assess the potential for affective interpersonal but remote haptic communication, we devised a three-part study (summarized in Fig 3.2), as well as a device validation pilot, in which we:

1. **Built an 8-tactor wearable prototype and conducted a small pilot on**

$N_{pilot} = 12$ people to evaluate the feasibility of haptic messaging for affect communication.

2. **Surveyed** $N_{survey} = 201$ **people about messages they send** to people they want to maintain touch relationships with despite obstacles such as distance or health issues. From this data, we constructed 10 scenarios that capture realistic context which might naturally prompt touch as a communicative element.
3. **Collected haptic message designs** by 10 dyads ($N_{design} = 20$ individuals) in close co-habitation relationships for their partners. These messages are contextualized by scenario prompts and include a personal wild card message of their choice.
4. **Collated interpretations** from $N_{interpret} = 21$ individuals who experienced haptic messages designed by strangers, their partners, and themselves a week earlier.

We assessed the physical aspects of the haptic message designs from (3) by intended emotion, identified features offering the greatest insight, then incorporated these into a machine learning model predicting emotion from message. We report how machine recognition of the emotional scenario prompts compared to that of human interpretations from (4), broken down by the interpreter’s relationship to designer: stranger, partner, or self (ordered in overall increasing interpretation rate). We describe the strategies that participants took in designing and interpreting encodings, noting where partner-focused strategies perform better than non-partnered counterparts; and suggest improvement priorities for the next iteration of a haptic messaging prototype. Finally, we observe how the interaction experience excited a spirit of play and whimsicality in design and recognition – an intuitive key in unlocking the privately shared tactile language between partners.

Overall, this paper contributes:

- a compilation of results from human and machine recognition of emotion-based intent in haptic messages;

- evidence that shared history influences the interpretation of playful affective haptic messages;
- a summary of design strategies and engineering parameters of haptic messages created by and for partners in close relationships, and a synthesis of our results into insights to inform future systems to support effective interpersonal haptic messaging.

3.2 Background

When studying machine-mediated haptic expressions of emotion, we want the touch to be representative of a genuine affective experience [39, 42]. In this work, we are further interested specifically in purposeful emotive interactions [148]. Thus, we present relevant related work and explain how it has informed our approach in two parts: (1) machine-mediated touch interaction where participants are (2) grounded in real emotions rooted in familiar events generating authentic touch expressions.

Machine-Mediated Touch and Display Expressivity: Defined as “the ability of one actor to touch another actor over a distance by means of tactile or kinesthetic feedback technology”, machine-mediation differs from direct touch where actors physically experience and reciprocate social touch in-person [107] (p153). There are many wearable or handheld devices (both research prototypes and commercial products) that purport to bridge physical distance to enable social touch [131]. The form of touch varies: the *Hey* bracelet¹ uses squeeze sensations (a motor rolls to tighten the band); *Shaker* [284] transmits a shake via a current between connected solenoids; the research tool, *The Tactile Emoticon System* is a glove form factor that transmits and receives pressure, heat, and vibration, concluding that interpretative value may hinge on message personalization [224]. Since studies have shown that users can infer nuanced affective information even from tuning simple vibrations alone [107, 249, 257], we inspect whether affective meaning can be made through custom designed vibrotactile messages, and examine the addition of making it a 2D travelling signal played out over time.

Haptic Spatial Animation to Leverage Expressive Sketching: Haptic anima-

tion [136] has demonstrated that perceptually interpolated (i.e., *animated*) spatial vibrotactile display creates an intriguingly varied and rich design space [137, 248]. Tactor arrays have been used to create interesting effects. From early attempts to discern simple directional lines [290], to simulations of real-world haptic experiences like a snake crawling up an arm [259] or a cat walking across one’s back [136], the field has made great strides in approximating convincing haptic effects from simple vibrations.

The density and positioning of the tactor array depends on the sensitivity of the body part stimulated. For instance, [82] developed a 3x3 tactor array that was sufficient for the entire back to feel fully activated; while [248] used a trapezoidal 5-tactor array embedded in a chair. A forearm is much more sensitive with two-point discrimination (the minimum distance where two distinct points can be differentiated) recorded at about 30.7-45.4mm from 43 subjects [205]), so devices need not be more dense than this linear distance. Distinct excitations were distinguishable only about 30-40% of time (chance 14.3%) at 25mm apart [54], suggesting that most vibrations from neighbouring tactors within this distance may be experienced as a continuous (i.e., without segmentation). At 22-44mm linear tactor distances, an illusion could be created of a snake moving in various ways across the arm [259]. We built our display to be within the two-point discrimination range [205], similar in range as [259].

The tactor array needs design tools for building haptic animations into socio-affective touch. Given that even relatively low-fidelity sensations can be emotively expressive [256], context may be as important to interpretation as the sensation itself [36]. Therefore, we explored the impact of context, and privilege a sender’s design experience over high fidelity display. Specifically, we anticipate that access to a spatiotemporal design palette will allow participants to define vibrotactile messages with greater personal significance with the potential for novel haptic experiences and expressions. We draw inspiration from the haptic design palette presented by [247] and affective vibrotactile parameters proposed by [249, 257], to develop an accompanying haptic sensation editor.

Emotion-Related Remote Touch Between Strangers: With direct human-to-

¹Descriptions available at <https://feelhey.com/pages/about>.

human touch established as a medium for emotion content [115], it follows to ask how much of this emotion encoding and decoding is retained when direct touch is intercepted by another medium or device. [11] had participants generate emotion-laden handshakes using a commercially available force-feedback joystick and found that those sensations were human-interpretable at roughly twice that of chance (33%, where chance was 1 in 7 or 14%).

Even without direct device contact, haptic sensations can communicate emotional content. The UltraHaptics system sends ultrasonic air pressure waves to deliver tactile sensations mid-air. [207] asked participants to design sensations that represent the emotions elicited by a provided picture, by modulating frequency, duration, and intensity. Another set of participants then rated how well suited some 10 haptic sensations were to a given picture. Again, there is evidence that mediated social touch can communicate emotion between people: participants consistently rated the haptic description designed with the picture with “high appropriateness”. Here we ask: since close relationships create more opportunity for communicating through touch, *what is the difference in emotion recognition between strangers vs. that of close partners?*

Relationship as Context: While most people would intuitively accept that humans can communicate emotion through touch, it is still somewhat surprising when sentiments like *anger*, *love*, *gratitude* can be recognized at rates above chance by strangers in a lab directly touching one another’s forearm [114]. Certain emotions have better recognition rates in the US vs in Spain; cultural relationship may explain some of these differences [114]. Studies on facial emotion recognition found that mutual cultural membership adds contextual background for how an expression may be made [118] which in turn influences emotion interpretation. Our interpretation of touch is similarly influenced [55] wherein culture defines who, when, where, and how we touch one another. So what happens where touch history extends beyond being simply cultural? Turns out that even in machine-mediated touch, relationship context (e.g., are we partners, friends, work colleagues?) is crucial for generating and interpreting Tactile Emoticons [224].

Many more factors contribute to the contextual framework that ultimately informs how a touch between two people is perceived [18]: the relationship between

them, the events triggering the touch contact, the environment and backdrop, and each participants' comfort with emotional expression. With respect to relation, Thompson *et al* examined touch interaction between couples and found that partners were better able to distinguish typically self-focused emotions like *embarrassment*, *envy*, *pride* than strangers [296]. Since technologically-mediated touch seems to follow common patterns for relationship contextualized haptic interaction [224], we wonder if *recognition improves with relationship closeness* where emotion-based touch messages authored by the participant themselves, their partner, and strangers may be successively less interpretable. Quantifying reasons for the difference in recognition rates influences how we structure our touch communication systems and interaction design.

3.3 Materials and Methods

A consumer-ready haptic messaging device would require careful iteration over hardware, functionality, and user experience. Here, we demonstrate a proof of concept for encoding emotional content into haptic animation – a necessary first step (process summary in Fig 3.2).

All experiments were conducted in 2019 and in accordance with the organization's ethical policy regarding human participant testing (with protocol as conducted later approved by WIRB, ref# AGHM-2019). Environmental Health and Safety approved the device prior to the study.

3.3.1 The Haptic Display

We describe the specifications of our haptic animation prototype and the interface for designing the sensations.

Building the Prototype

Our prototype's custom haptic display features eight voice-coil vibrotactile tactors (model: TEAX13C02, Tectonic Elements, UK²) which are arranged in equilateral triangles (35 mm sides) along two columns, and padded with laser-cut Polyureth-

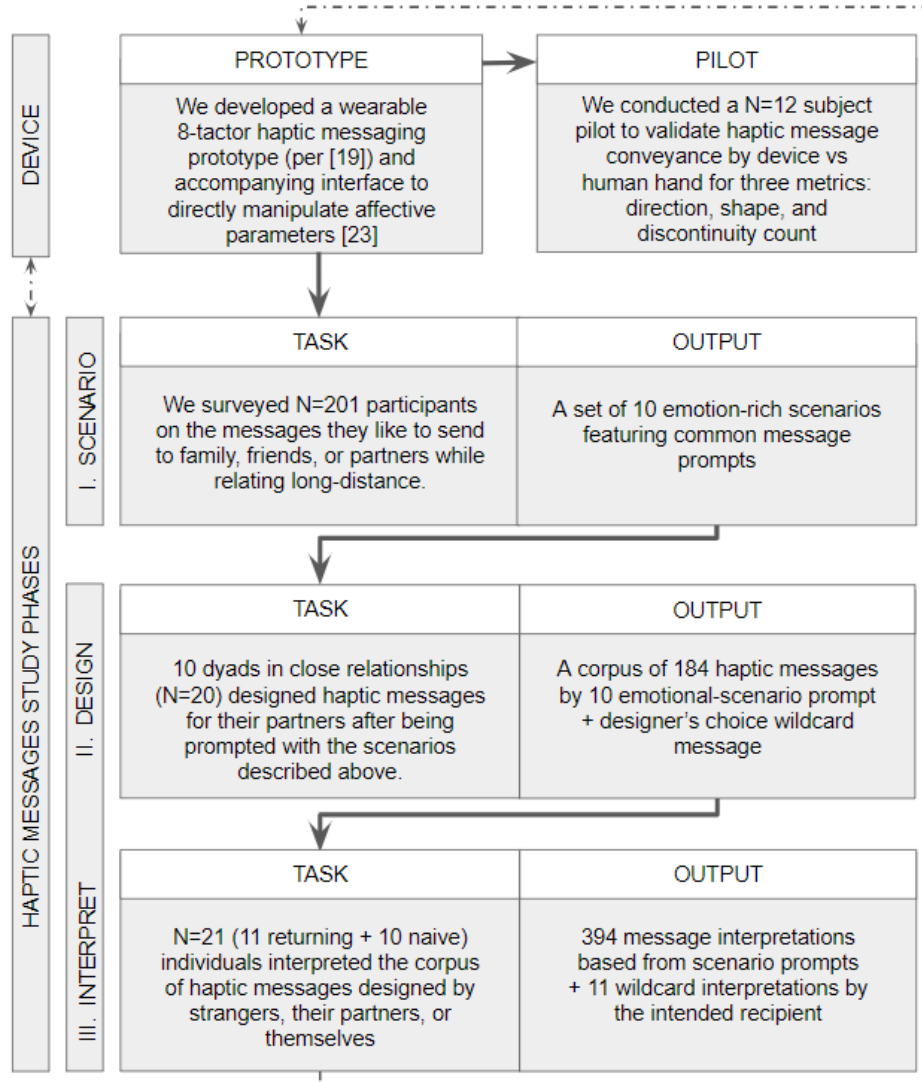


Figure 3.2: An iterative process (dashed lines) of developing a device suitable for a haptic messaging application. For this paper (solid arrows), we built and piloted a wearable device and conducted a 3-phase haptic messaging study based on designing and interpreting haptic sensations rooted in emotion-laden scenarios commonly experienced by members of close long-distance relationships.

ane foam (Figure 3.1(b)). Each tactor is housed in a 3D-printed casing to isolate electrical components from directly touching the skin, and covered with an insulated cover. The vibrating element of each actuator is covered with a thin 15mm diameter disk that contacts the skin. The actuators are computer-controlled using an audio interface (Motu, USA, model 24Ao³) and powered with a set of audio amplifiers (MAX98306⁴).

Tactor Position

The inner forearm from wrist to elbow is tactually sensitive, socially discrete, convenient and practical, without hindering the hand [50, 199]. These traits make it an excellent candidate for placing a tactile display. To leverage the wide design space of haptic animation [136, 247], we followed [248]’s blueprint to create a medium-fidelity prototype of a haptic animation display with tactors positioned as vertices of equilateral triangles but with intertactor distance scaled down to 35mm to be wearable on the arm (rather than embedded in a chair-back as in [248]). This distance was so chosen to fall within the two-point discrimination range of 30.7mm - 45.4mm for the forearm [205] in order to render a continuous vibration sensation between adjacent tactors (Figure 3.1(b) shows relative positioning). When engaged in the frequency range of 20-300Hz, the tactor array on the device has a conservative active surface area that covers the contact area of the device at about 60mm wide by 155mm long (since each tactor has a two-point discrimination radius of 30mm). Strapped tightly to an adult forearm, the device casing that houses the tactors has a height of 20mm over the contact area.

Message Editor & Process

So that lay designers could access this prototype function with minimal learning, we developed a rudimentary graphical user interface (GUI) in which a designer can define their haptic sensations by manipulating a set of vibration parameters (summarized in Table 3.1) and drawing directly on the representative display area

²Material specifications for each element can be found at: TEAX13C02, Tectonic Elements: <https://www.tectonicaudiolabs.com/product/teax13c02-8rh/>.

³MOTU 24Ao: <https://motu.com/products/avb/24ai-24ao>.

⁴MAX98306: <https://www.maximintegrated.com/en/products/analog/audio/MAX98306.html>.

of a touch screen (Figure 3.1(a)).

In playback, the resulting sensation is graphically presented by a circle that follows a designer-laid track (visible only to the designing partner), while playing out only tactually on the message recipient's arm in the same timeline.

Terminology: We refer to the touch screen region as the *drawing surface* where designers can define the size of the circle or *brush* to draw a *track* defining the path that the haptic sensation travels. A track can consist of one or more *strokes* which are continuous drawn segments. The circular brush's diameter represents how wide the track feels – e.g., a large brush radius signifies a wider or thicker track line such that tactors passing under the brush are activated. Upon playback, participants feel the haptic *animation* which is the sensation of the recorded design. A *haptic encoding* refers to a tactile animation designed to communicate a specified intent; this haptic signal together with the intent are a *haptic message*.

Parameters editable through the GUI (Table 3.1): As well as spatial path and dynamics, the designer can modify vibration intensity by editing the signal waveform's *amplitude* and *frequency* (both defined before drawing a message that does not vary over the course of a single message). *Brush size* refers to the circle diameter (i.e., width of the drawn track), such that tactors within the brush's circular boundary are activated at varying intensities based on *diffusion type*. Three *diffusion types* (linear, quadratic, exponential) allow users to define how sharp and focused the animation brush feels across the path's breadth during tactile replay. With linear diffusion, there is a gradual fade from the ball's centre to its border; exponential causes the most dramatic fade with most of the sensation near the path centerline; quadratic is in between.

Drawing process: The drawn track defined the x-y coordinate of the 2D drawing surface as well as the time variation of the stroke. After drawing a track, users can then record their haptic message, play it back, and edit the sensation using all available parameters including speeding up or slowing down the animation along the existing track. Any edits to track placement requires a new design.

Our priority was for designers to achieve their haptic-message vision. Although we tried to make the GUI usable, its success was not a focus at this stage, so a researcher helped designers navigate the interface.

Table 3.1: Editable parameters for haptic message design

Parameter	Controls	Range/Options
track	message motion	unspecified limit
travel rate	cursor velocity	unspecified limit
brush size	effector area	0 mm - 165 mm
diffusion factor	effector sharpness	0 mm - 165 mm
diffusion type	effector gradient	linear, quadratic, exponential
amplitude	effector strength	defined by amplifier volume & tactor specs
frequency	vibration	0 Hz - 500 Hz
duration	message length	unspecified limit

Understanding the Display: A Pilot Study

To ensure that participants could tactually perceive the overall physical sensations rendered by our haptic display (a prerequisite of interpreting their intended meaning), we conducted a pilot in which we asked participants to re-draw eight researcher-defined haptic animations, so chosen to vary shape, area covered, segment count, direction, curvature and angles.

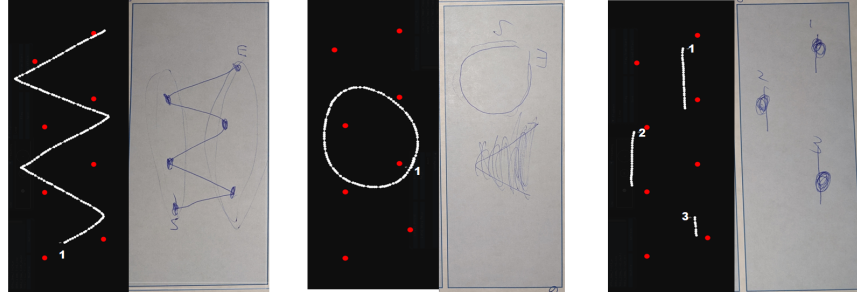


Figure 3.3: Representative perceptibility pilot results. A comparison of the test design (L) and a participant-drawn interpretation (R) from each of the three images above (chosen from the 8 message trajectories as depicted in Table 3.2). Each continuous segment is labelled with the order of its playback.

To reduce novelty effects of the device, participants who received the stimuli from the haptic display were first introduced to the prototype in a sandbox session




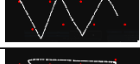



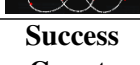
where they could experiment with the controls and sensations. We asked them to find and tell us the parameter ranges on brush size, diffusion, amplitude and frequency where haptic sensation felt both clear and comfortable; then set participant-specific ranges based on their preferences. In general, pilot participants found the sensation most pleasant at low vibrational frequencies, reporting a range of $\mu(\sigma)$ of 31.7 Hz (11.7) to 91.2 Hz(25.6).

To assess device rendering accuracy as compared to actual continuous touch contact, $N = 6$ people wore the device to receive the eight researcher-defined haptic animations, and as a control, another distinct group of $N = 6$ people had a researcher draw the same eight shapes with an index finger, counterbalancing order. All participants received the stimuli on their non-dominant arm and used their dominant hand to draw out the track they were feeling on their forearm. They were instructed to use arrows or mark (S)tart and (E)nd points to indicate direction as well as order all discontinuous segments (Examples in Figure 3.3). We assessed exact agreement (no partial credit) between participant drawing and original design on three metrics: (1) discontinuous segment count; (2) direction of the motion; and (3) the shape of the 2D track.

Outcomes

As summarized in Table 3.2, participants correctly distinguished the segment count (device $\mu = 6.7$, $\sigma = 0.5$; human $\mu = 7.5$, $\sigma = 0.8$) and direction (device $\mu = 7.3$, $\sigma = 0.8$; human $\mu = 6.5$, $\sigma = 0.5$) for at least 6 of 8 animations. They were less successful at recognizing the exact shape defined by the track (device $\mu = 4.7$, $\sigma = 0.5$; human $\mu = 5$; $\sigma = 0.9$). The triangle shape was hardest to decipher with the angles often drawn as discontinuities. Due to the small sample sizes, we ran separate Kruskal-Wallis tests (KW-test for non-parametric comparison between group measures) to compare success rates of recognizing each of segments, direction, and shape as drawn by human touch (control) vs device stimuli (as in Table 3.2). Results showed no significant differences ($p > 0.15$) across all three factors. We then calculated the effect size using epsilon squared (compatible with the KW test), obtaining very weak effect sizes at all measures at $\epsilon^2 < 0.003$. Therefore, we proceeded with this device iteration assuming that the animation display and GUI are

Table 3.2: Pilot results: Test Stimulus Perception. Values are the number of pilot participants able to draw the exact (no partial credit) number of Segments, Direction of motion, and overall track Shape of the test sensation as delivered by the haptic device (Dev) or control human researcher (Hu). Each participant group contained $N = 6$ individuals for 48 trials (8 test stimuli x 6 people).

Test	✓Segmen	✓Directio	✓Shape
	Dev — Hu	Dev — Hu	Dev — Hu
	4 – 6	5 – 6	3 – 3
	6 – 6	6 – 6	3 – 3
	6 – 5	6 – 5	6 – 5
	6 – 6	6 – 5	4 – 5
	3 – 4	4 – 4	2 – 3
	5 – 6	5 – 3	3 – 4
	4 – 6	6 – 5	3 – 3
	6 – 6	6 – 5	4 – 4
Success Count:	40 – 45	44 – 39	28 – 30
Device Success:	83.3%	91.7	58.3%
Human Success:	93.8%	81.3%	62.5%

basically adequate for the purposes of this study. To prepare for a learning curve, later study phases incorporated time to get used to the device and a sensitivity test to ensure that design and interpretation participants could discern the stimuli with similar success as in the pilot.

3.3.2 Message Meaning Phase: MTurk Online Survey

To build a realistic set of message meanings (i.e., situated within scenarios) for which participants would design descriptive haptic encodings, we surveyed indi-

viduals who had been in at least one long-distance relationship ($N = 201$, Amazon Mechanical Turk). They answered questions about the kinds of messages they did or would have wanted to send within a prominent relationship that had established a high-level of interpersonal touch prior to being long-distance.

Survey respondents (mean age 33.4 years) reported on relationships with a spouse or romantic partner (68.2%), parent (12.9%), friend (10.9%), child (3.5%), grandparent (2.5%) and sibling (2.0%), most of which (54.5%) were long-distance for at least a year. We compiled their free-form responses to “*What message would you most like to send to your loved one?*” (all of which were conversation initiations). Two independent raters looked for the main themes of the intended ensuing conversation; resolving the two independent lists, we agreed on the 8 categories listed in Table 3.3. The most common category involved bids for conversation without a specific topic, including general updates that often open with *Hey*. Next are a series of emotions elicited: *excitement*, *miss you* (sometimes also referred to as *longing*), *sadness*, *love*, *anger*, *gratitude*, *anxiety*. We found that some messages expressed high urgency or arousal, through use of all-caps (e.g., “*IS SOMETHING WRONG?!?*”) so we noted the extremes as *calm*, *attention*. For each category (plus *calm* and *attention*), we created scenario prompts to ensure common contexts with implicit roots in emotion. Finally, since each partner pair has a distinct communication style and background, we added a *wildcard* message scenario for a total of 11 prompts for message generation listed in Table 3.4.

3.3.3 Design Phase: In-Person Dyads

We recruited ten dyads (self-reported as 9 male, 11 female; aged $\mu = 29.5$ years, $\sigma = 4.7$) who happened to be of diverse cultural backgrounds from the US, UK, Belgium, Ecuador, Russia, China, Thailand, and India, that together were representative of the population of the greater Seattle area and who reported being in comfortable touch relationships. Participants generated haptic messages based on the scenario prompts in Table 3.4. Nine of the dyads were in committed long-term romantic relationships, and one in a best-friendship (relationship length $\mu = 6.9$ years, $\sigma = 4.9$ years). Each pair was living together at the time of the study. All individuals reported a high level of comfort with electronic and messaging devices.

Table 3.3: Eight categories of crowdsourced messages to send to loved ones from a survey of people ($N = 201$) in long-distance relationships

Type	Count	Sample Message
Update (Hey)	96	"Hey how's your day going?"
Excited to Visit	65	"I can't wait to see you so we can celebrate!"
Miss You (Longing)	41	"I'm thinking about you and I miss you so much."
Sadness	34	"Can we Facetime soon? I'm just disappointed right now."
Love	26	"I've been wanting to say ' <i>I love you</i> ' but just didn't know how."
Anger / Frustration	18	"It's so frustrating to be apart so long."
Grateful for Relationship	17	"I'm so glad we met. Thank you for being so patient."
Anxious for Well-being	15	"How are things with your family? I'm worried about the kids."

Sessions took ~ 60 minutes, and each participant was compensated with a small honorarium of \$75 USD.

Familiarization

After getting comfortable with the device and controls, participants proceeded to perceiving and distinguishing a useful range of sensations. To reduce novelty effects and establish a message sending / receiving experience, dyads were first given a chance to play with the device together. Specifically, we allotted time for: (a) *Sandbox mode* to establish a messaging context and reduce novelty effects, consisting of 10-15 min of playing with the device and sending instant messages, wherein Person A draws on the touch screen while Person B wears the display and vice versa; (b) a *sensitivity test* so researchers can ensure participants are able to perceive basic encoding elements; (c) a *learning phase* for designers to familiarize themselves with the GUI and the device capability.

One participant wore the haptic display device on their forearm while the other drew on the touch screen interface to transmit real-time messages. They switched

Table 3.4: Scenario Prompts for Haptic Message Design and the Number (#) of Designs for Each. Participants designed at most 1 encoding per scenario (some ran out of time before completing all 10 prompts).

Code	#	Scenario Prompt
anger	17	Your partner has done something careless that has set off a sequence of inconveniences and you are frustrated that this has happened. Again.
anxious	17	You're really nervous about a big presentation you have to make today and you are asking your partner to reassure you or wish you luck.
attention	19	You need to talk to your partner RIGHT NOW. It is an EMERGENCY.
calm	15	You've just had a very relaxing massage and you think your partner should try one too.
excited	16	You've just received the big news you were hoping for! You want your partner to know that you want to celebrate together!
gratitude	17	You found out that your partner has done something really kind for you unexpectedly. You want to thank them and let them know that you appreciate them.
hey	16	You want to see if your partner wants to connect when they have a moment.
love	16	You want to send a message that assures your partner, confirms that you value them and reciprocate their feelings about you.
miss	17	You are folding laundry and you are remembering the day you met your partner and want to tell them you're thinking of them.
sad	17	You are really sorry for what happened and you want your partner to know.
wildcard	17	Are there any other kinds of messages that you would like to send to your partner? Please design one of your choice.

roles halfway through the sandbox session; we gently suggested a switch around the 5-min mark but allowed dyads autonomy to decide for themselves. While we didn't originally set out to analyze engagement, we noted dyads' light teasing and giggling at each other's interpretations during the sandbox mode.

To evaluate sensitivity, researchers played three of the *haptic tracks* from the device validation pilot (a triangle, a circle drawn counterclockwise, three dashed

lines) for the device wearer; their partner also came up with a few of their own designs. After each haptic track, we asked them to describe the direction, segment count, and overall shape (same parameters selected from the evaluation pilot). We had decided in advance to omit participant designs created by any individual unable to correctly describe two of the three parameters of each of these basic shapes. All participants exceeded this sensitivity threshold.

Participant Message Creation

To design the haptic encodings, each dyad member was paired with a researcher and led to separate locations to work independently. Researchers helped their participant use the interface to ease the learning curve. Within a session, each participant designed encodings for up to 10 scenarios chosen from Table 3.4 in random order, ending with a *wildcard* message of their choice. Participants were able to edit, save, and playback their encoding until satisfied before progressing to the next scenario. To ensure that the haptic designs were aligned with the scenario’s intended sentiment, researchers read the scenario aloud and asked participants to describe the kinds of feelings that the scenario incited for them before proceeding with the design process.

Production of Encodings for Interpretation Phase

From this message generation phase, we retained only instances of the haptic encodings where participants verbalized an emotion language that agreed with the sense of the scenario prompt (up to 10 plus a *wildcard* per participant). A total of 184 unique haptic message encodings were used for the next phase: 167 from the predefined prompts plus 17 *wildcards* (see Table 3.4 and Figure 3.4 for counts and duration respectively).

3.3.4 Interpretation Phase: In-Person Singles

To assess how well these haptic encodings could be understood, one week after the Design Study phase we invited all message generating participants back to ‘receive’ a set of messages. Of the original 20, 11 returned from the design phase, including four partner pairs; the other nine were unable to return for the followup

due to scheduling constraints so we recruited 10 naive participants. $N = 21$ individual participants were played a set of designs “as if they had been sent from close friends or family”. Though no specific sender was specified, returning participants were informed that the message set would include some of the messages their partner had designed for them. Sessions in this study phase took about 30 minutes total where participants were again given up to 10 minutes in sandbox mode first for familiarization.

Encoding Set

Returning participants were given a set of 20 haptic encodings, carefully selected to contain two of each scenario prompt where encodings contained (in random order) messages made by themselves, their partner, or a stranger (on average, 5.2, 5.7, 7.9 messages respectively). For the 10 naive participants, all interpreted encodings that were made by a stranger ($\mu = 18.7$ messages each). Participants worked through as many haptic encodings as they could within a 30-minute block (up to a maximum of 20). They were given a list of the 10 original design phase scenario prompts (see codes in Table 3.4) and were asked to match the scenario to their interpretation of the haptic encoding.

Procedure

Message recipients were not able to see the graphical message track at any time during the interpretation phase; they could only feel the sensations. Participants were first asked to freely identify their first impressions of the sensation or what they would instinctively assume the intent to be if they had received the message. They were encouraged to elaborate in a think-aloud format for each haptic message. If unsure, they could replay a messages without limit and/or skip to send it to the back of their queue. After experiencing the message, participants were asked to associate each of *1-very likely*; *2-somewhat likely*; and *3-very unlikely* tags with at most one of the 10 scenarios using a Qualtrics survey application (offered to eliminate the pressure of a single forced choice). They could elect to leave a tag unattached. For interpretation classification, we used the scenario tagged by the highest likelihood (i.e., if *1-very likely* was unattached, we use *2-somewhat likely*

as their top choice). Messages that only get a *3-very unlikely* tag was treated as uninterpreted.

Finally, all repeat partner-dyad members were asked to interpret the wildcard message that their partner designed specifically for them. Eight (of 11) returned to the study explicitly for this purpose. “*I moved an appointment for this! I’m excited to feel what he created for me.*” - P04b

3.4 Analysis & Results

Our analysis was guided by two primary questions.

(1) Are there feature subsets so characteristic of certain classes of emotion scenario that they may be machine distinguishable? That is, how much of the sentiment can be described by the physical engineering parameters alone?

We report recognition rates generated using Random Forest (RF), a popular technique for machine learning of affective touch interaction [42, 87], particularly where low data density or strict computation limits preclude sophisticated deep learning models (the former being relevant here).

(2) How well, comparatively, can *people* recognize the sentiment behind the messages if the haptic sensation was designed by (1) a stranger, (2) their partner, or (3) themselves a week ago? For this, we viewed results of our machine classification alongside each of these human-recognition situations.

Analysis consisted of (Section 3.4.1) haptic encoding feature extraction and consideration by emotion scenario; (3.4.2) qualitative examination of the designs themselves, aimed at understanding the diversity of approaches taken; and (3.4.3) a quantitative look at machine and human interpretation accuracy.

3.4.1 Features and Parameter Analysis

In the following, we describe the data preparation and machine classification process for determining message intent. In overview, we carried out feature extraction for each message, then evaluated feature significance to identify parameters that have the most impact on distinguishing scenarios.

We extracted 82 features (summarized in Table 3.5) from each of the 167 haptic messages. Outside of the user-defined track drawing parameters (*diffusion factor*

Table 3.5: Summary of Features Extracted. Parameters expressed in **bold-type** are vectors – scalars otherwise.

Stat Function	Parameter	Count
-	DiffusionFactor, DiffusionType, BrushSize, Frequency, TotalSegments, Duration, AreaDisplacement, AreaDistance, TrackDistance	9
max, total	displacement, velocity	4
auc	x / time, y / time, y / x	3
min, max, mean, median, var, auc	x, y, speed, angle	24
(min, max, mean, median, var, auc) of set of segments	duration, distance, max and total displacement, speed, angle, area	42
Total feature count:		82

and *type*, *brush size*, and *frequency*), we included track characteristics such as the number of discontinuous *segments* and *duration* (s) of the entire message (see Fig. 3.4 for duration distribution by scenario prompt). We also calculated two kinds of displacement: *max displacement* refers to the Euclidean distance between the maximum and minimum coordinates for x and y values of the track; and *total displacement* refers to the Euclidean distance between starting and end points of the drawn track. *Max* and *total velocity* use the respective displacement values over the elapsed time. The three Area Under the Curve (AUC) calculations are based on y values by x, x over time, and y over time. We also calculated the full *Track Distance* as the distance travelled within the display area (excluding any discontinuous jumps); *Area Distance* is the smallest rectangular area bounded by the track and *Area Displacement* is the smallest rectangular area bounded by the start and end points of the track. We calculated a set of six statistical functions (*min*, *max*, *mean*, *median*, *variance*, and *auc*) for all remaining vectors: the full set of x and y coordinates across the haptic display as well as the **speed** = $\frac{\Delta(x,y)}{\Delta t}$ and **angle** $\theta = \tan^{-1} \frac{\Delta y}{\Delta x}$. Finally, we created **duration**, **max displacement**, **total**

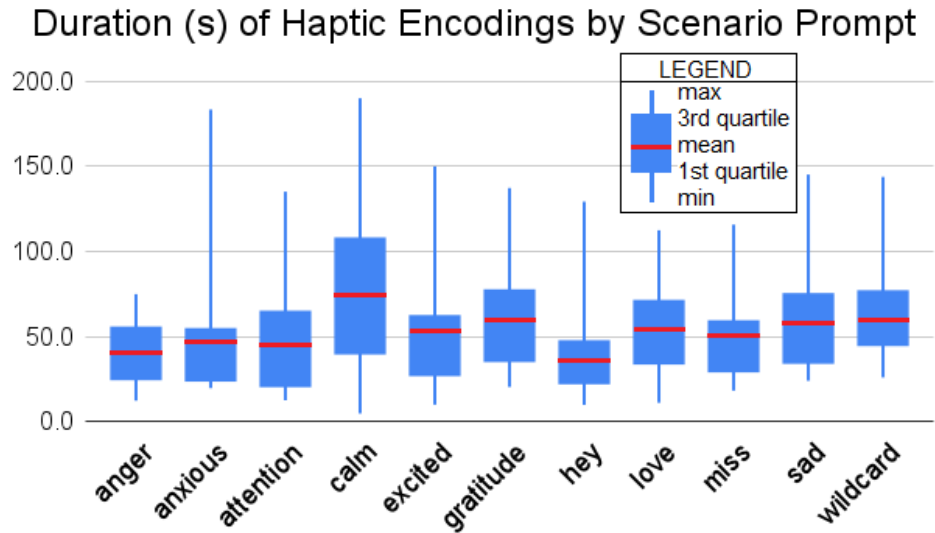


Figure 3.4: Participants designed messages of unspecified duration where **calm** has the largest variation in duration and **anger** the shortest.

displacement, distance, speed, angle, displacement area vectors comprising the disconnected segments of a message and calculate the same six statistical functions for each parameter.

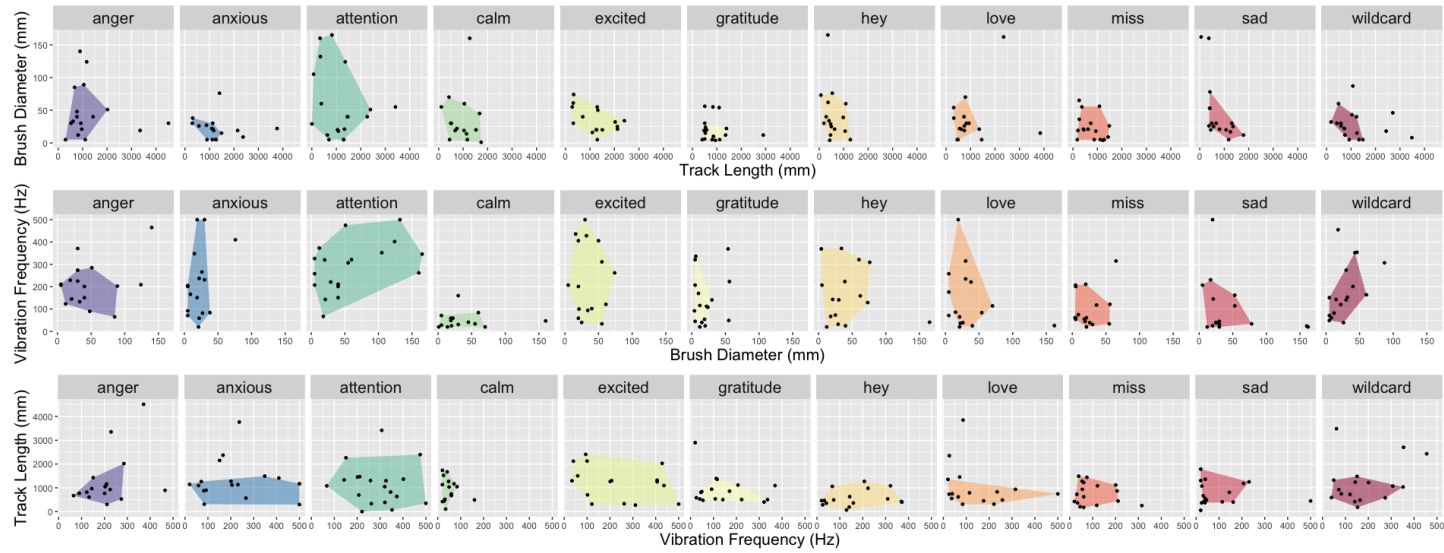


Figure 3.5: Designer-created Message Parameters of Track Length, Brush Diameter, and Vibration Frequency for each Emotion Prompt (cross-reference by emotion word for prompt in Table 3.4), including the wild-card message which participants created for their partners. Here, we see that *calm* tends to small and slow designs with small brush size and track lengths and low vibration frequency; in contrast, *attention* has a large range of vibration frequencies and brush sizes though mostly small to medium track lengths.

To see how physical parameters correlate with the sentiment intent in message generation, we ran a series of ANOVAs on all 82 features. For the features directly controlled by the message designer, specifically *Diffusion Factor*, *Diffusion Type*, *Brush Size**, *Frequency**, *Segment Count*, *Duration*, *Track Distance**, *Area Displacement* and *Area Distance*, the three marked with * were significant at $p < 0.05$. We plotted these dimensions (Figure 3.5) to get a sense for how distinct the characteristics are from the messages generated in each emotion-laden scenario.



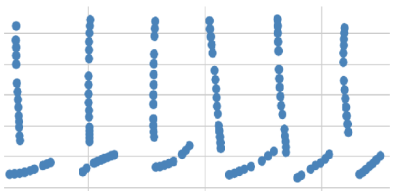
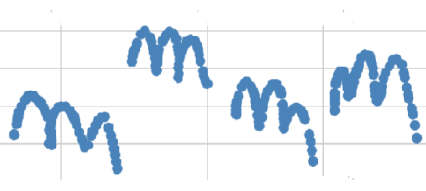
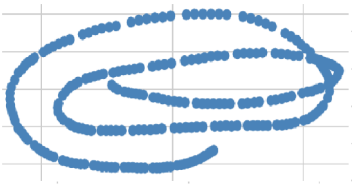
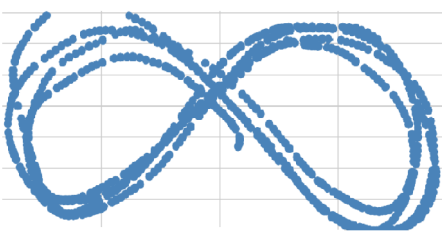
3.4.2 Qualitative Analysis of Participant Designs

Even in the Sandbox (where partners swapped messages face-to-face), distinct approaches emerged for creating track-dependent haptic messages, and continued to develop during the sessions. To capture the diversity of these approaches, we performed a thematic analysis on the 17 designs wherein three raters independently determined 3-6 groupings of the messages. After a lengthy discussion, all raters converged on three high-level categories such that all wildcard designs fall under at least one of (1) direct transcription of some visual representation (either a drawing or writing or other symbology); (2) a rhythmic repetitive sensation that leverages temporal patterns; and (3) distinctive physical sensations that exploit the contrasts between continuous/discontinuous or sharp angular/soft fluttery (see Table 3.6).

Spanning these categories, we observed approaches that varied both in form of expression (e.g., spatial vs. temporal patterns) and in drawing on private shared context, generic references, or abstractions (e.g., literally spelling with letters). These approaches reappear in the unconstrained wildcard messages, which may emulate real world use.

We expected spatio-physical sensations designed to evoke interesting haptic experiences. However, we were intrigued to see participants like P10a make pictograms that visually represented the message intent to be traced out on the recipients' arm in a haptic message (Table 3.6). Similarly, P05b wrote out a word in Simplified Chinese. We also note the surprising interplay between repetitive discontinuous segments to play with temporal patterns, drawing more on rhythms than spatial representation.

Table 3.6: Three raters determined that all 17 wildcard message design strategies fell in three categories, with illustrative examples.

Examples from Categories of Design Strategies	
Transcription of Visual Representation	
p10a - wanna grab coffee?	p5b - write chinese character; playful
	
Temporal Patterns	
p06a - I have a story to tell you!	p09a - let's eat I'm hungry
	
Spatio-Physical Sensations	
p07a - I'm happy	p04a - check out this cool feeling!
	

3.4.3 Interpretation Accuracy

When people touch one another, a plethora of social cues and emotional content can be conveyed, particularly between intimate partners [193]; considerable affective content is also communicated through touch between strangers [114]. We wonder if social content communicated in close relationships can be sent and interpreted more accurately compared to that between strangers, particularly when we add a machine interlocutor. For insight, we first compared recognition accuracy by machine and human strangers; neither of these have personalized training nor

shared history with message designers.

Figure 3.6 uses accuracy confusion matrices to compare classification outcomes for four cases of interest: by machine, human stranger, partner, and self. Correctly classified instances are on the diagonal.

By Machine Stranger: We consider the use case where a machine interprets the message and communicates a best guess to the intended recipient. For this to work, we conceive of a procedure where a model is trained on the haptic encodings labelled with the emotion ascribed to the presented scenario. We selected Random Forest (RF) as our classifier, as the literature has shown RF to work well with affective and social touch [41, 42, 87, 96, 152]. We found (Figure 3.6(b)) that 10-fold cross validation using a subject-dependent (touches from same participant may be in both training and test data) RF classifier on 167 messages achieved an overall accuracy of 18.6% (chance 10%).

By Human Stranger: Machine classification and stranger interpretation are both performed on messages by unknown designers and there is little or no shared history, so it is interesting to compare these results. When we asked participants to evaluate designs by strangers, out of 274 trials (in which some of the 167 encodings are repeated), people’s best guess – the prompt they thought was the most likely match – was accurate 17.9% (chance 10%) of the time, compared to 18.6% by the RF classifier.

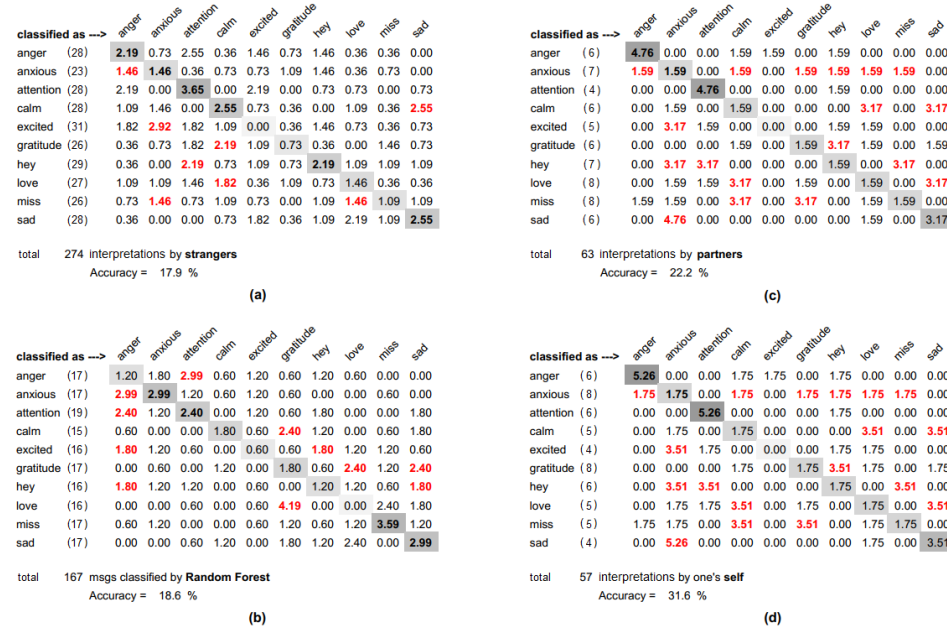


Figure 3.6: Confusion Matrices Comparing Interpretation Accuracy of Affective Content for each Haptic Message (count of interpretation instances). In order of increasing accuracy, by (a) human strangers, (b) machine stranger (Random forest classification), (c) designer's partner, and (d) the designer themselves, a week later; chance = 10%. **Red** values indicate where the highest mis-classification rate matches or exceeds the diagonal.

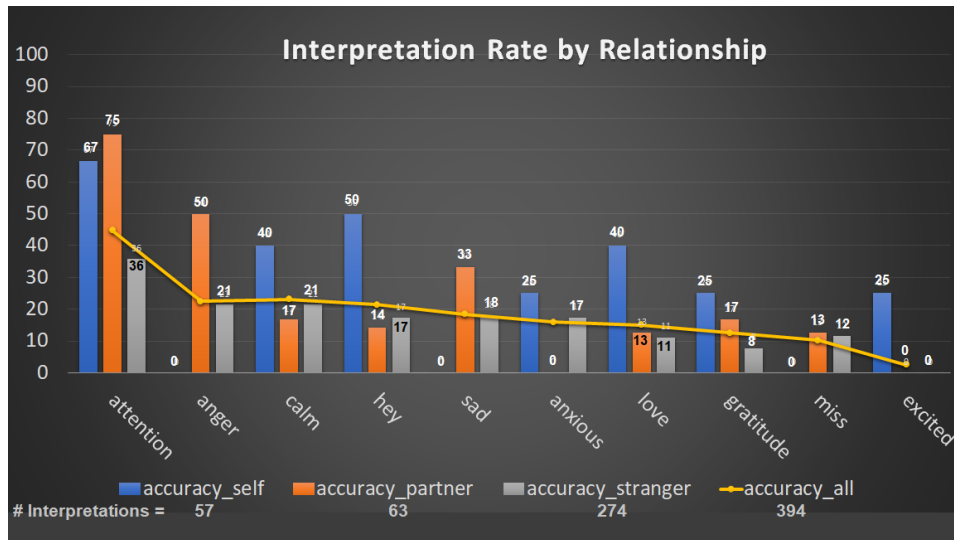


Figure 3.7: Interpretation Accuracy (%) by Message and Relationship, ordered by decreasing overall recognition accuracy.

By Relationship: We can look at recognition rate of the intended emotion in each message prompt to see how sentiment communication varies depending on the relationship (through the confusion matrices of Figure 3.6). Here we see that overall, message designers recognize their own messages most often (Fig 3.6-d, 31.6%) and that of strangers (Fig 3.6-a, 17.9%) the least, with partners in between (Fig 3.6-b, 22.2%).

However, the story becomes more complicated: some prompts defy the expectation that interpretation accuracy increases with relationship closeness. We summarize message interpretation accuracy by relationship in Figure 3.7. Interestingly, *anger*, *miss*, and *sad* are consistently poorly recognized by the designers themselves.

Table 3.7: Wildcard Messages designed for and interpreted by partners. Participants designed one wildcard message each.

Designer's Intent	Recipient's Interpretation	Got it
where u, wanna chat	My partner wants to talk to me when I'm available. She is probably in a good mood and maybe has some good news to share	Y
A kiss!	xxx (kisses)	Y
I'm heading home. I love you. Can't wait to see you.	I think it means love; kindness; reassuring and feels really massage-like	Y
I'm happy	happy lovely	Y
shrimp. because he took more than i did even though I made dinner	is she angry about all the shrimp I ate?	Y
we fit together. (drew a puzzle piece)	I think this means connection. like a jigsaw puzzle	Y
eyeroll. there's a person who annoys us and is currently annoying me	oh yeah. she's irritated. This is the pacing that she'd use to say "oh. my. GOD." with	Y
kisses; like affection in a playful way	!!? What is wrong?	N
check out this cool feeling	repeated figure 8; if he was exceedingly silly with the message, I think this could be "death to Videodrome. Long live the new flesh!"	N
Hello	It feels like writing? But I can't tell what it is):	N
Let's eat. I'm hungry	Don't really know what it is, but it's something positive?	N

Of Wildcard Messages: Of the 20 design phase participants, 11 returned for interpretation. We played each one the wildcard message that their partner designed specifically for them. These messages were completely open-ended (presumably very low chance of randomly guessing correctly).

For wildcards, we marked an interpretation *Correct* when two independent scorers agreed that the interpretation matched the intent. Scorers looked for word matches, synonyms, and common sentiments. Interestingly, 7 of 11 message recipients were able to correctly interpret the message (summarized in Table 3.7), a recognition rate of 63.6%, higher than most other emotion prompts.

Some interesting interpretations include P07a’s design jigsaw puzzle pieces (Figure 3.1(d)) to communicate that they “*fit together*”. Partner P07b recognized the message as representing “*connection ... like a jigsaw puzzle*” (Table 3.7).

A more abstract design is P06b’s signal where two lines followed by a lengthy swirl communicates an “*eyeroll*” about “... *a person who annoys [my partner and me]*”(Fig 3.1(e)). Upon feeling the encoding, partner P06a immediately recognized it as communicating irritation since the rhythmic pattern was reminiscent of “... *the pacing [P06b would] use to say ‘oh. my. GAWWD.’ with*” - P06a.

3.5 Discussion

Our primary study goal was to learn how to more effectively leverage social touch in haptic messaging. While our results generally corroborate the literature asserting that contextual background and shared history play an important role in the perception of emotional content [21, 36, 296], we now return to our research questions and discuss how our data provides evidence toward answering them.

3.5.1 Message Design Observations

Our study’s encoding designers were tasked with communicating rather complex social meanings. They were given only a short time to learn an unfamiliar device limited to sensations that are low-resolution and unnatural relative to direct human touch, albeit shown to encode interpretable affective content [146, 256, 257]. Furthermore, participants designed with researcher assistance which may introduce other biases – e.g., participants may accept researcher suggestions more readily

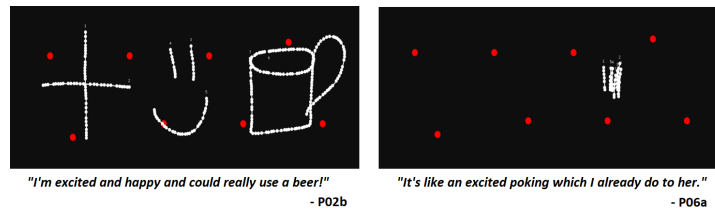


Figure 3.8: Two very distinct ways of designing for the same *excited* message prompt.

than if they were alone. Departures from the ideal use aside, we wish to know more about the haptic message design experience: how do people approach haptic message design given our current prototype and scenario prompts?

Do Design Strategies Reveal Shared History? By examining the wildcard message designs, we suspect shared history is embedded in the visual drawing strategy and the rhythmic/temporal patterns, particularly evident in P06a recognizing her partner's idiosyncratic speaking cadence (Figure 3.1(e)).

The shape-drawing strategies also evoke common backgrounds, illustrated by P5b's communication in their native written language, who shared that "*this is like a game that my parents used to play with me as a child. They would write a character on my arm or back and ask me to guess what they wrote*". P07b's design based on connecting puzzle pieces (Figure 3.1(d) – interpreted accurately despite its complexity) made us wonder if this pair might enjoy doing jigsaw puzzles together.

While acknowledging that the touchscreen interface may have suggested a visual approach (drawing and writing) we are encouraged by the range of content that screen-sketching supports. Partners appear to draw recognizable patterns from multiple senses – most notably visually and vocally (as in Figure 3.1(e)) – raising questions about how these and other approaches might evolve with more time.

Are there Universal Messages? Designers played with physical parameters like frequency and brush diameter to determine whether the sensation communicated the right intent. Figure 3.5 highlights variation in the interaction between significant design parameters, creating a parameter "footprint" by message. Designers

seemed to match lower arousal [239] message intentions (like *calm*, *miss*, *sad*) using small brush diameter and low frequency.

Interestingly, *attention* was often designed with the largest brush at the highest frequencies, but also had the largest footprint – the greatest design variation across all designed messages. This suggests that there are other factors attributable to highly human-interpretable messages which transcend design consistency: i.e., some concepts might be broadly amenable to many representations, or they might be extremely personal and our participants were able to find the particular encoding that worked for their partner.

3.5.2 Interpretation Rate

We preface discussion of accuracy by noting that we carefully built on lessons from past research, and while it is not useful to compare results directly due to divergence of our evaluation objectives in this novel application space, we consider important takeaways from our work and relate it to literature where similarities exist. For example, we acknowledge that many works have very distinct types of touch constraints, instructions, and evaluation methods, yet it is interesting to note that in some cases, comparing with other device-mediated affective touch (e.g. [11]), we may still see proportionally similar performance (accuracy that roughly doubles chance recognition). In all cases, mapping vibrotactile sensations to emotive thoughts can be non-intuitive, and in many ways it is remarkable that interpretation accuracy would ever exceed chance.

Our work and others' demonstrate that digitally mediated affective communication is feasible. However, it is important to keep in mind that these relatively low instance counts (particularly for the self and partner interpretation sets as seen in Figure 3.6) provide only limited insight into message efficacy. Thus, our primary benchmark of comparison for this exploration is what people themselves can do in the same study conditions.

Why do Designers Not Recognize their Own Designs? Generally, relationship closeness does influence interpretation accuracy of social touch ([193] and here, Figure 3.7). Thus we would expect designers to be best at recognizing the messages they had designed. However, this is not always the case. Figure 3.7 highlights how

message prompts for *anger*, *miss* and *sadness* or *sorry* seem to be recognized more accurately by partners and strangers than by their designers. While this certainly needs more investigation, we observe that touch behaviours communicating these sentiments are especially likely to be directed to another – e.g., one is unlikely to miss, or show longing for, oneself. Possibly, this leads to our being less likely to recognize our own touch when expressing sentiments with this quality.

How Can Wildcard Recognition Be So High? While most messages have interpretation accuracy under 35%⁵ (chance 10%), wildcard messages – with no specified prompt and thus, no fixed interpretation option – is recognized surprisingly well at 7 correct interpretations out of 11 messages (Table 3.7). The *wildcard* messages may be the best examples of closeness in relationship improving interpretation of message intent. People with a shared history can draw from a wealth of experiences to generate creative messages, even idiosyncrasies from other modalities. Speech rhythm and cadence (P06a) is one example, but we can imagine haptic sensations that emulate an impatient tapping foot or short strokes that channel dumbfounded cartoon blinks.

Shared *recent* context surely impacts message interpretation rate. If P09a and P09b had been fighting on the way to the study session about one eating more than their share of dinner, then shrimp may have been on both of their minds (Table 3.7). A recent charged memory could make a highly specific message easy to read, maybe even irrespective of the haptic design. Removing the shared history by getting strangers to interpret the general content *wildcard* messages (excepting *disagreement over shrimp*) could be illuminating.

What is the Potential for Machine Recognition? Machine recognition rates were comparable to stranger recognition (18.6% and 17.9% respectively, chance 10%). Comparing Figure 3.6(a) and (c)’s confusion matrix diagonals shows that messages of *anxious*, *gratitude*, *miss* and *sad* are better recognized by machines than human strangers. Figure 3.5 shows distinct patterns of common design parameters, particularly for *anxious* (small active area with small brush diameter and low track length across a large vibration frequency range) and *gratitude* (similarly small active area but with vibration frequency mostly in the low end). Perhaps these ranges are stat-

⁵Notable exceptions are *attention* (~75%) and *anger* (~50%).

istically distinct but tactually imperceptible, making it more difficult for human interpreters.

Because message designers created only one design for each message prompt, we have a sparse training set with no repetition on the interaction of two important dimensions (designer and message). Affective touch interaction is individual, so machine recognition increases dramatically with more person-specific training [42]. We see an opportunity for additional training samples to complement shared contextual history where machine recognition may serve to support interpersonal message interpretation accuracy.

3.5.3 The Messaging Experience

Despite not setting out to evaluate the ‘fun factor’ of the messaging experience on our haptic animation prototype, we discovered that the design sessions where pairs worked together (sandbox mode) were often punctuated with giggles and gentle ribbing (“*What? You mean you can’t feel that’s a heart?!*” – P02b to P02a) as close friends and partners were first developing a sense for how to use the device. We noted that at least 11 of 20 participants spoke about playfulness, happiness, and/or laughter while designing their *wildcard* messages, revealing extra pleasure in imagining their partners puzzling out meanings involving private context and some amount of effort. The fun generally emerged through affectionate collaboration – it was not a solo activity. Here, we discuss valuable observations of the design-interpretation process and speculate about improvements necessary before it can become a viable communication channel.

How did Individuals Vary Across the Pipeline of Interpretation? We expect that successful haptic messaging likely depends on both the subjective tactile perception and interpretation of the message intent; each of these are themselves complex processes. The first depends on display performance (nature of stimuli, resolution, dynamic range, etc), skill of the designer’s use of it, and the individual perceptual sensitivity of the recipient. The second is where we were able to focus more in the present study, looking at factors like relationship and message type.

Our study protocols included checking on comfort and threshold proficiency in using the system, but did not comprehensively measure individual acuity or its

components or demographic influences. Thus we cannot speak to the degree to which perceptual challenges (as well as stimulus type suitability – i.e., of vibrotactile modality for affective messages) impacted individuals and pairs' ability to use the system and enjoy the interactions.

We informally observed a wide range of skill in both individuals' and dyads' ability to construct or fully carry out a communication chain – typical for the studies involving either tactile acuity or emotional intelligence. Acuity arises both from sensitivity to ranges of sensation, and at a higher level, the ability to mentally integrate then identify shapes that are received as spatio-temporal line drawings on the skin. For example, we were particularly impressed by the recipient of the wildcard jigsaw puzzle pieces: these pictographs are complex with many vertices and two separate but closely set, compatibly interlocking components. This integrative feat seems remarkable, and likely beyond the capability of most other participants or indeed the researchers. However, it is a fascinating example of what might be possible, and may have been aided by contextual factors that improved this recipient's guessing odds.



Figure 3.9: P02b particularly enjoyed designing haptic messages after a first try on *anger*, and imagines developing a vocabulary.

What is the Longitudinal Prognosis? These were one-shot design efforts. People can learn to adjust to a partner and to a communication medium. We wonder how individual and dyad performance would improve over time, and how pairs might evolve and enrich their communication style – what strategies they would come to rely on or discard; how memory would work, stability of vocabulary (already

anticipated by P02b – Figure 3.9), what kinds of context (short or long term) they would leverage when given the opportunity. We wonder if the interaction would become more engaging and/or valuable as a core communication modality when more familiar, or soon set aside. Our present results are a promising start, but real answers await longer studies and a device and editor that could function in everyday life.

How Could We Improve our Prototype? Inherent to the messaging experience is the device and interface. We built a minimally viable prototype to establish the feasibility of affect-content communication via a wearable haptic animation display. Our findings in design variation, interpretation rate, and overall enthusiastic reception suggest that even our simple, low-resolution prototype can open up a rich and evocative haptic playground. To further enrich the experience, subsequent iterations of the hardware could integrate smaller, more powerful tactors to increase end effector density (i.e., allow for more tactors to fit in the same surface area), which may afford more intricate designs. The most apparent example of spatial resolution or sensitivity discrepancy was evident during the Sandbox phase with people like P02b incredulous with her partner’s (P02a) inability to recognize the more intricate shapes. *“It’s clearly got angles though babe!” – P02b when P02a mis-identified an octagon as a circle.*

The design interface could also be amended with more fine-grained control mechanisms. Although only one design participant out of 20 asked, we can envision a scenario where experienced users may want to design messages with time-varying frequency and amplitude, dramatically increasing the range and complexity of the design space.

Corroborating findings from the hand-based Tactile Emoticon (featuring haptic sensations of temperature, vibration, and pressure) [224], we posit that so long as users are provided a sufficient customization range for designs and design strategies, partners may play around to come up with something that works for them, regardless of device sophistication. We expect iterations of device and messaging application to inform one another; here, we present a promising proof of concept as a strong starting point.

3.6 Conclusions

We presented a multi-phase study on machine-mediated social touch to shed light on how people might create and interpret emotion-encoded haptic messages. We used a custom wearable spatial tactile display, and an interface for participants to compose spatiotemporal patterns. The study’s scope included scenario prompt sourcing, message encoding design, and message interpretation. Its design and analysis highlight the influence of relationship and shared context on how communication plays out. We summarize the key findings sparking future lines of inquiry.

1. A shared history between message designer and interpreter generally improves message comprehension: private inside jokes are a great strategy; individuals are not always great at reading their own tactile-writing. Overall, message interpretation accuracy increases from strangers (17.9%) to partners (22.2%) to message designers themselves after a week (31.6%). However, partners could understand 7/ 11 of open-ended *wildcard* messages, a surprisingly high accuracy given their unconstrained content. We posit that shared contextual knowledge is of great value; and further note that the sharing was almost always both humorous and private in nature – couples sharing a private inside joke, with their common experience the key that unlocked understanding.

We also found that some messages were poorly recognized by designers themselves compared to their partners. We speculate that physical manifestations of anger and missing or longing are not often directed at ourselves, so we are less likely to recognize our own – but need more than one message per designer to be sure.

2. Machines are about as good as strangers at haptic message interpretation (for now). Our machine classifier recognized message intent with 18.6% accuracy, comparable to that of strangers (chance 10% in all cases) where closer relationships between sender and receiver serve to improve interpretation rates overall. We imagine that insofar as machine-‘strangers’ and human-strangers lack shared context, both relationships are similarly distant with the message sender. Future work could inspect whether personalized training may offer a machine analogue for ‘history’.

3. Individual design strategies may co-opt other modalities. Some designers produced visually recognizable pictograms to communicate message sentiment –

puzzle pieces and happy faces drawn on the touchscreen – while others played with rhythm – discrete taps simulating excited poking behaviour, or strokes and spirals timed to mimic an idiosyncratic speaking cadence. Given this early diversity, how might design behaviour mature if pairs had more time to trade messages?

Next Steps: This study has benchmarked interpretability rates and highlighted encoding strategies for a relatively expressive haptic display (relying on spatiotemporal animation, supporting drawing-type designs). Looking ahead, our findings underscore the importance of considering design strategies when choosing displays and editing systems, that maximize expressive capability; and that dyad communication is highly unique, rich with many characteristics helpful in maintaining emotional connection in relationships.

Obvious next steps are to develop physical displays that are practical and comfortable in real settings yet at least as expressive as the one used here – and are fun to use. Then it will be possible to launch studies that monitor how vocabulary used by dyads develops and enriches or withers over time, and the contribution this kind of communication makes to pairs who cannot be together. **So when is a haptic message like an inside joke?** We think it matters only when there's someone you care to share it with.

Chapter 4

Machine as Emotion Influence: An Investigation into Machine Breathing as a Fear Contagion

Summary

People often physically cling to others when afraid [114, 115]. This response occurs for good reason, as physical touch can downregulate negative emotional experiences [58]. However, touch might be ineffective—or even detrimental—for downregulating fear experiences if the others being touched are experiencing and expressing fear themselves. We posit that touching others expressing fear can guide perceptions of fear via the detection of distinct respiratory patterns, which might cause emotion contagion and consequently bolster rather than inhibit one’s own fear response. To test this hypothesis, we built plush robots with motorized plastic ribcages that were manipulated to contract and expand to simulate human breathing patterns. We asked participants to hold these robots as we measured their heart rate (HR) before, during, and after watching a fear-elicitation stimulus. Consistent with our hypothesis, participants who interacted with robots that exhibited fearful breathing patterns perceived the robots to be fearful and experienced a pronounced increase in their own HR, compared to participants who held stable- breathing and non-breathing robots. These results suggest that touching or clinging to others to downregulate one’s own fear may be detrimental if the other is also displaying bod-

ily movements and physiology of fear. This study is the first to test whether distinct artificially generated respiratory patterns influence human emotion contagion via touch, and to do so by measuring human emotions through autonomic nervous system activity.

Significance Statement

Physically touching others during anxiety-inducing events can downregulate one's own fear experiences. Yet, touching others might be ineffective—or even detrimental—for downregulating fear if the others being touched are also expressing fear. Using plush robots with motorized ribcages that were manipulated to contract and expand to simulate human breathing patterns, we found that participants who held fearfully “breathing” robots showed in increased heart rate compared to those who held calmly “breathing” or static robots. Results thus suggest that human autonomic nervous system activity is influenced by emotion contagion occurring through touch, and individuals should therefore use caution when seeking to downregulate their emotions by touching pets, support animals, and possibly other humans who are confronting the same event.

4.1 Introduction and Background

People often touch or even cling to others when they are afraid. A frightened child might grasp a parent when startled, and adults will grab partners or friends during scary movies [114, 115]. There is good reason for these behaviors; the mere presence of others can help downregulate negative emotion, and interpersonal emotion regulation benefits are heightened by physical touch with humans [58, 67, 231, 323] and service animals [191]. In certain situations, however, relying on physical touch to downregulate fear might backfire. When others touched are also experiencing and expressing fear, touching or holding them allows for the felt perception of their fear [114, 115], which could, in turn, bolster one's own fear response through a process of emotion contagion. Emotion contagion occurs when a person “catches” or comes to feel the same emotion as expressed by someone else [111, 300]. A large body of research has demonstrated that emotion contagion can occur by visually observing facial and postural expressions of emotion;

observers come to feel or express the same emotion themselves [65, 216, 300]. In real-life situations of fear, however, contagion may be less likely to occur through the observation of others' visible expressions, because fearful individuals tend to focus their attention towards the fear-eliciting stimulus rather than other interactants [185, 209]. Nonetheless, even without visual attention directed towards a fear-experiencer, fear may be communicated and become contagious via touch. When individuals feel fear, they display rapid and deep breathing (e.g., hyperventilation), an observable pattern that is different from that which occurs during low arousal emotions such as sadness or calmness—emotions instead characterized by slower and stable breathing [28, 220]. This physiological indicator is not specific to humans; many animals – including cats and dogs, which are commonly used for emotional support – exhibit changes in their breathing when frightened [128, 129, 213]. Given that breathing requires expansion and contraction of the chest, alongside other discernable body movements, individuals might accurately perceive others' fear through touch if their touch allows them to detect the bodily changes that occur with respiration. Supporting this expectation, medical professionals are encouraged to both look and feel for evidence of chest movements (expansion and contraction) to establish breathing during clinical assessments [154, 226]. Breath patterns may therefore constitute an effective and widely generalizable mechanism for communicating emotion through touch (across species), and, because emotion perception can elicit emotion experience via emotion contagion, observing others' breathing patterns through touch may influence the emotions experienced by interactants. Few studies have tested whether distinctive breathing patterns elicit perceptions of distinct emotions or cause emotion contagion. Although studies have demonstrated that robots mimicking mammalian breathing patterns shape observers' perceptions of robots' emotion and likeability [35, 157, 293], these studies have not assessed participants' own emotion experiences, so it remains unclear whether emotion contagion can occur from touching a robot exhibiting artificial breathing. Other studies have found that individuals interacting with a robot exhibiting movements designed to mimic calm mammalian breathing patterns report feelings of calmness and stress reduction [10, 192, 253]. However, these studies did not test whether different breathing patterns presented by robots— for example, fearful versus calm patterns – have different effects on emotion contagion. Further-

more, past research on this topic has been limited by small sample sizes affording low statistical power and reduced generalizability ($Ns < 38$), and has relied heavily on within-subject manipulations that increase participants' awareness of manipulated changes to robots' apparent breathing patterns, thus increasing demand characteristics (i.e., participants may become aware of the experimenter's hope that they will respond differently in different conditions). Overall, prior research suggests that breathing patterns effectively communicate diagnostic information about fear experiences when these patterns are visually observed or felt, and humans seek to touch or hold others as a means of downregulating their own fear. It remains unclear, however, whether touching or holding others who display a variety of breathing patterns differentially influences individuals' own emotional or physiological experience. More specifically, previous studies have not addressed the question of whether a frightened human is likely to experience fear inhibition or enhancement when touching another individual who displays rapid, seemingly fearful breathing.

4.1.1 The Current Research

We tested whether: (a) humans can detect and recognize “fear” by touching a robot displaying chest movements simulating hyperventilation, and (b) touching a robot displaying these movements enhances humans' own fear experiences. To address these questions, we built a robot with a motorized plastic ribcage, such that we could manipulate its precise “breathing” patterns to simulate fear and calmness. We recruited participants to hold this robot while sitting still and watching a series of videos; participants' heart rate was measured throughout the entire procedure, including before, during, and after presentation of a fear-elicitation video stimulus. We hypothesized that participants holding a robot that displayed a fearful breathing pattern seemingly in response to a fear-eliciting stimulus would detect and interpret the robot's movements as conveying fear, and would demonstrate an increase in their own heart rate, compared to participants holding a robot displaying stable breathing or no breathing movement. Other non-verbal expressions of fear may also communicate and evoke fear responses in participants, particularly fear expressing sounds like gasping, heavy breathing, and other vocal expressions.

To mitigate robot audio expectations and minimize any mechanical sounds, participants wore noise cancelling headphones to listen to the video stimulus' audio. This research is the first to manipulate artificial respiratory patterns of an organism interacting physically with human participants, and to test for human emotion and ANS contagion via touch.

4.2 Methods

4.2.1 Participants

One hundred and seven undergraduate students from the University of British Columbia were recruited to participate, but we excluded four individuals whose heart rate data could not be matched to their self-report data due to a technical error. Our final sample thus consisted of 103 undergraduates (73% women, 26% men, 1% other; 49% East Asian, 20% White, 8% Middle Eastern, 6% Hispanic/Latino, 2% African American, 15% other; $M_{age} = 20.59$ years, $SD_{age} = 2.93$ years). A post-hoc power analysis indicated that this sample size provided greater than 99% power to detect the observed change in HR within the fearful breathing condition.

4.2.2 Procedure

All participants watched an identical series of video clips totalling 288 seconds while their heart rate was monitored and they held a fur-covered robot. Participants were instructed to hold the robot in their arms, hugging it against their chest (i.e., as they might a close relationship partner, parent, child, or pet), generating maximal physical contact. Participants kept their right hand under and left hand on top of the robot, with a PulseSensor heart-beat detector on the middle finger of their right hand. They wore a pair of Koss UR23IK headphones to deliver sound accompanying the video clips and minimize disruption from incidental mechanical noise from the robot. Participants were instructed to avoid engaging in any excess movement to prevent interference with the heart-beat reading from the finger sensor. Participants were also instructed to watch the computer screen throughout the duration of the experiment (see Figure 4.1 for experimental set up).

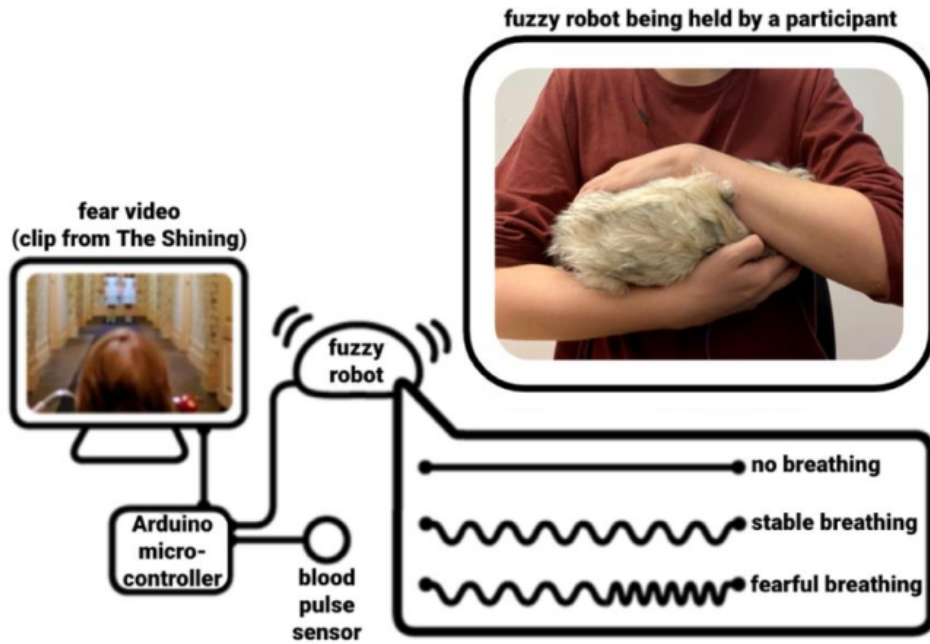


Figure 4.1: The robot structure (top right) and a diagram of the participant experience of watching a fear validated video while holding a fur-covered robot that demonstrated one of three breathing patterns (bottom right) manipulated between participants.

The robot was roughly the size of a small house cat. It had a soft, plush, and furry covering (see Figure 4.1). We designed it to be shaped and sized like a small pet instead of a human, for several reasons. First, intimate and convincing physical contact between humans, like clinging or hugging behavior, occurs following a high threshold of complex interpersonal, cultural, and social norms [91, 287]; in contrast, humans approach and touch domesticated animals with a much lower threshold [133]. Second, it is considerably more feasible to simulate the appearance of a furry animal-like robot than a human, and this simulation is crucial because robots that approach human-like appearance but do not achieve it can elicit unsettling discomfort – an effect called the “uncanny valley” [88, 260]. Finally, furry zoomorphic robots displaying breathing motions have previously been validated as reliably communicating emotional content [34, 35, 254, 322].

To acclimate participants to the robot, they were asked to sit in a chair while

holding and examining the robot. When each participant was ready to begin the experimental procedure, they were set up with headphones and the heart-beat sensor worn on their finger. Participants next watched 114 seconds of videos that were intended to acclimate them to the experimental context without eliciting strong emotions. The first 30 seconds consisted of a black screen accompanied by no sound, followed by an 84-second video of a snail crossing a wooden plank, which was accompanied by ambient sounds of nature in the background. The snail video was found on YouTube, where it was labeled “The most boring video in the world. The snail”. After the acclimation period, participants watched another 30-second black screen, followed by an 84-second fear-elicitation video clip taken from the movie *The Shining*; this clip has been used in prior work, and rigorously validated to elicit the distinct emotional experience of fear [103]. Following the fear-elicitation video, participants viewed a final 60-second black screen. The 20 seconds of black screen directly preceding and following the fear clip constituted our pre-elicitation and post-elicitation measurements of heart rate (respectively). However, we also planned to construct Locally Estimated Scatterplot Smoothing (LOESS) lines with 95% confidence intervals to measure and visualize changes in HR continuously throughout the procedure, given the high likelihood of uncovering non-linear changes in participants’ HR. Finally, all participants completed an online questionnaire before being debriefed. While watching all video clips, participants were randomly assigned to hold the robot while it demonstrated one of three breathing patterns, manipulated between participants: no breathing, stable (calm) breathing, and fearful breathing. In the no-breathing condition, the robot showed no movement throughout the entire session (i.e., from the beginning of the first black screen of the acclimation period through the last second of the final black screen after the fear-elicitation clip). In the stable breathing condition, the robot displayed a stable expansive and contractive movement throughout the session, designed to mimic a chest cavity when breathing at a rate roughly equivalent to human resting respiration (i.e., 14 breaths per minute; BPM). In the fearful-breathing condition, the robot engaged in a breathing pattern with modulated acceleration. This began with chest movements identical to those in the stable breathing condition (14 breaths per minute), which occurred for 144 seconds: throughout the 114s acclimation period and 30 seconds of black screen preceding the fear-elicitation video.

Over the course of the fear elicitation video, these movements changed to accelerate the expansion/contraction rate up to 30 cycles per minute (30 BPM). This acceleration was designed to simulate fast breathing and hyperventilation consistent with human fear experiences [28, 220]. When the fear-elicitation clip ended, the robot’s movements were decelerated, and after 60 seconds its apparent breathing rate returned to the pre-elicitation stable pace, which was maintained until the conclusion of the session. Figure 4.2 shows a visualization of the breathing patterns conveyed by the robot in each condition, along with the order of the videoclips shown to participants.

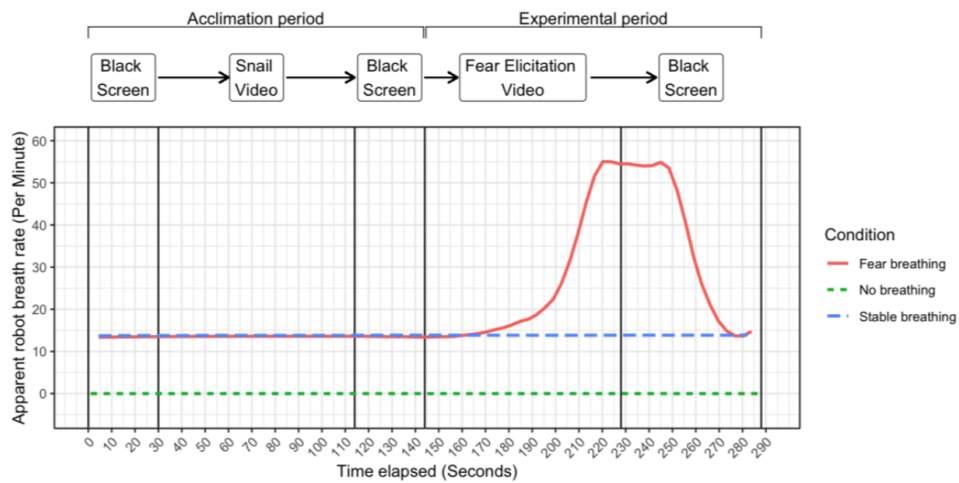


Figure 4.2: Visualization of the order of videos presented during the procedure and corresponding robot breathing via the manipulation of simple symmetric sine waves throughout the 4 minute and 48 second procedure. The apparent breathing rate in the fear condition plateaued between 220 seconds and 245 seconds, due to mechanical limitations of the robot motor prohibiting it from moving at a faster pace.

The no-breathing condition was intended to function as an inactive control (e.g., baseline) condition, reminiscent of experiencing fear while engaged with something akin to a stuffed animal. The stable-breathing condition functioned as an active control – ensuring that any effects in the fearful breathing condition were not attributable to the general presence of movement suggesting life. By including multiple robot “breathing” conditions and a non-breathing condition, this design

allowed us to test whether the specific movement pattern displayed by the robot in the fearful-breathing condition – and not the robot itself, or the appearance of breathing alone – upregulated participants’ fearful emotional response. Following the human-robot interaction, all participants completed an online survey consisting of self-report measures asking them retrospectively evaluate the robot’s behavior and their own emotions throughout the experiment. This survey was completed up to five minutes after the conclusion of the human-robot interaction. Prior to conducting the study, we did not know how long any emotional effects of the videos and robot interaction would linger, and although we endeavored to capture condition-based differences in state-level emotions after the conclusion of the task, and included these measures as exploratory dependent variables, we also suspected that any subjectively experienced emotion might have dissipated after five minutes.

4.2.3 Materials

Robot Construction

To develop the robot prototype, we followed social robot design for single degree-of-freedom motion [34] and, using a similar template structure, created the wishbone template to form the robot’s skeletal structure. We laser cut the wishbone shape in varying sizes so that, even under a thick fur cover, the back of the robot had a ridged spine-like feel. The main form was comprised of two parallel panels; each was comprised of a long and narrow piece with a large round bulb at one end, much like a tomahawk steak. When the two panels were lined up in parallel, the bulb portion formed the head (where a central motor was housed) and the long curved narrow pieces formed a track with notches in which to secure 16 wishbone-shaped pieces. The curvature of the wishbone sides formed ‘ribs’ and, by attaching strips of flexible plastic (23-gauge polyethylene) to the bottom of each set of ribs, we created a curved and lightly pressure-resistant soft robot ‘belly’, particularly evocative once the entire body was covered in a soft furry fabric. Fishing line was used to thread through each of the plastic strips of the belly and connect it to the central motor secured in the head. To build and manipulate breathing behaviors,

we used an Arduino Uno microcontroller to manage the motor. At the motor arm's maximum position, the fishing line pulls on the belly strips to simulate an exhale contraction; at the motor's minimum position, the fishing line is relaxed and the compliant belly relaxes similarly to express an inhale belly extrusion.

Participant Heart Rate

We assessed participants' heart rate (HR) throughout the procedure using a plug-and-play optical pulse sensor for Arduino. We chose to focus on changes in HR as our main dependent variable based on meta-analytic evidence that HR increases to a significantly greater degree during fear experiences compared to neutral (control), sadness, surprise, anger, and disgust experiences (i.e., all emotions compared to fear experiences in a meta-analysis by Cacioppo, Berntson, Klein, & Poehlmann, 1998). We used Kubios HRV Premium to convert the raw optical voltage data into R-R intervals (the distance between peaks in a sinusoidal waveform). Heart rate in beats per minute (BPM) was obtained via an arithmetic conversion.

Self-report Measures

Fear-elicitation manipulation check: Participants were asked to retrospectively recall how “Angry”, “Sad”, “Happy”, “Afraid”, “Surprised”, and “Bored” they felt while viewing the video of the snail and then while viewing the fear-elicitation video. For each of the two video clips, participants provided a rating ranging from 1 (Not at all) to 5 (Completely), for all six emotion prompts, for a total of 12 ratings per participant.

Robot Expression: Participants were asked to retrospectively rate the extent to which they perceived the robot to feel “Angry”, “Sad”, “Happy”, “Afraid”, “Surprised”, “Bored”, and neutral (“The robot did not feel anything”) while viewing the video of the snail, and then while viewing the fear-elicitation video. Participants provided separate responses characterizing the robot's feelings for the two events using a 5-point rating scale ranging from 1 (Not at all) to 5 (Completely).

Self-Reported State Emotion: After concluding the experimental session, participants rated their own current feelings on the state-level Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegan, 1988; 20-items), the Cur-

rent Mood Questionnaire (a measure of positive and negative valence and arousal; Feldman-Barrett & Russell, 1998; 12 items), and fear (Harmon-Jones, Bastian, & Harmon-Jones, 2016; 3-items). For all measures, participants responded using a 5-point rating scale, with higher numbers indicating more intense emotional experience.

4.3 Results

4.3.1 Manipulation Checks

Fear-elicitation video: To determine whether the fear-elicitation video successfully elicited fear, we compared participants' self-reported feelings of fear in response to the fear-elicitation video versus the snail video, using constructed multilevel models predicting self-reported fear from video type (snail versus fear-elicitation), along with random intercepts for participants to account for repeated-measures ($ICC = .08$). Results showed greater self-reported fear during the fear-elicitation video, $Beta = 1.41$, $t(102.93) = 14.96$, $p < .001$.

Robot Breathing: To determine the efficacy of our between-subjects manipulation, we tested whether participants in the fearful-breathing condition perceived the robot to be more afraid compared to participants in the stable-breathing and no-breathing conditions. Supporting the validity of our manipulation, participants indicated that the robot was more afraid in the fearful-breathing compared to the no-breathing, $Beta = -1.20$, $t(101) = -6.11$, $p < .001$, and stable-breathing conditions, $Beta = -1.19$, $t(101) = -6.00$, $p < .001$. There was no difference between the no-breathing and stable-breathing conditions, $Beta = .01$, $t(101) = 0.06$, $p = .95$.

4.3.2 Main Analyses: Do humans show emotion contagion from artificially breathing robots, via touch?

To test whether the robot's breathing pattern affected participants' physiological responses to the fear-eliciting stimulus, we conducted between-subject analyses to examine participants' HR during the 20 seconds directly preceding the fear-elicitation video ("pre-elicitation"; seconds 124 to 144 in Figures 2 and 3), and 20-seconds immediately after the video ("post-elicitation"; seconds 228 to 248 in Fig-

ures 2 and 3). We constructed a multilevel model predicting HR from robot breathing condition (fearful, stable, or no-breathing, dummy coded with fearful-breathing as the reference group), time segment (pre-elicitation versus post- elicitation), and condition by time-segment interactions, along with random intercepts for participants (ICC = 0.39). Participants in the fearful-breathing condition demonstrated a significant change in HR post-elicitation when compared to pre-elicitation, $Beta = .31$, 95%CI: [.24 to .37], $t(5494.41) = 9.02$, $p < .001$. Participants in the no-breathing condition demonstrated only a very small increase in HR post-elicitation compared to pre-elicitation, $Beta = .09$, 95%CI: [.02 to .16], $t(5492.94) = 2.52$, $p = .01$; significantly smaller than the change in HR observed in the fearful-breathing condition, $Beta = -.22$, 95%CI: [-.31 to -.12], $t(5493.63) = -4.32$, $p < .001$. Finally, participants in the calm breathing condition showed no significant change in HR post-elicitation when compared to pre-elicitation (although the effect was trending in the same direction), $Beta = .07$, 95%CI: [.00 to .14], $t(5492.53) = 1.89$, $p = .06$; significantly smaller than the change observed in the fearful-breathing condition, $Beta = -.24$, 95%CI: [-.34 to -.14], $t(5493.40) = -4.75$, $p < .001$, and no different than the change observed in the no-breathing condition, $Beta = -.02$, 95%CI: [-.12 to .08], $t(5493.74) = 0.44$, $p = .66$.

Together, these results suggest that participants in the fearful-breathing condition experienced an increase in HR between pre- and-post-elicitation, whereas participants in the no-breathing condition experienced a significantly weaker but still statistically detectable increase in HR, and participants in the stable-breathing condition experienced a still weaker increase in HR that was not statistically significant. Figure 4.3 shows a visualization of HR over time using Locally Estimated Scatterplot Smoothing (LOESS) with a span of .65, along with 95% CIs around LOESS lines.

To test the robustness of these results, additional models were constructed, resulting in similar patterns. Specifically, we constructed models including participant gender as a covariate, and removing participants who recognized the movie scene used in the fear-elicitation stimulus ($N_{final} = 79$). We also constructed models with HR centered around participants' personal baseline (with baseline defined by the average HR during the first 10 seconds of black screen preceding fear-elicitation stimulus); in one model baseline was included and in one it was as a covariate.

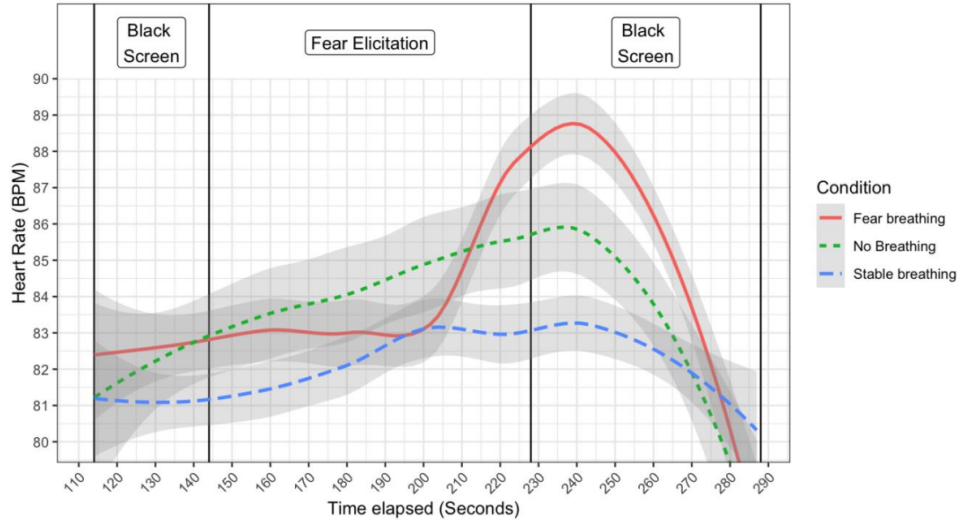


Figure 4.3: Physical setup showing room layout and relative positioning for participant and researcher over all stages of data collection.

In all four models with paired t-tests, participants in the fearful-breathing condition experienced an increase in their HR between pre- and post-elicitation ($Betas > .31$, $ps < .001$), whereas those in the no-breathing condition experienced only a small change in HR, ($Betas < .10$, $ps < .011$), which in all cases was significantly smaller than the change observed in the fearful-breathing condition ($Betas < -.22$, $p < .001$). Finally, in all four models participants in the stable-breathing condition demonstrated no significant change in HR between pre- and post-elicitation ($Betas < .08$, $ps > .056$). For full reporting of all models, see SOM.¹

We next tested the effect of robot-breathing condition on state-level self-reported emotion (ANOVA), which was collected at the end of the study. No differences emerged between the three breathing conditions for self-reported fear, $F(2,101) = 1.25$, $p = .29$, negative affect, $F(2,101) = 1.45$, $p = .24$, pleasantness, $F(2,101) = 2.49$, $p = .09$, unpleasantness, $F(2,101) = 0.90$, $p = .41$, high activation, $F(2,101) = 0.46$, $p = .63$, or low activation, $F(2,101) = 0.66$, $p = .52$. However, there was

¹Given that the robot's breathing rate did not change in the stable-breathing and no-breathing conditions, we cannot test for synchronization (i.e., with no variance in breathing rate, we cannot test for covariance with participants' HR, or differences in these relationships across conditions). For a visualization of participants' HR alongside the robot's breathing pattern, see Figure 4.4.

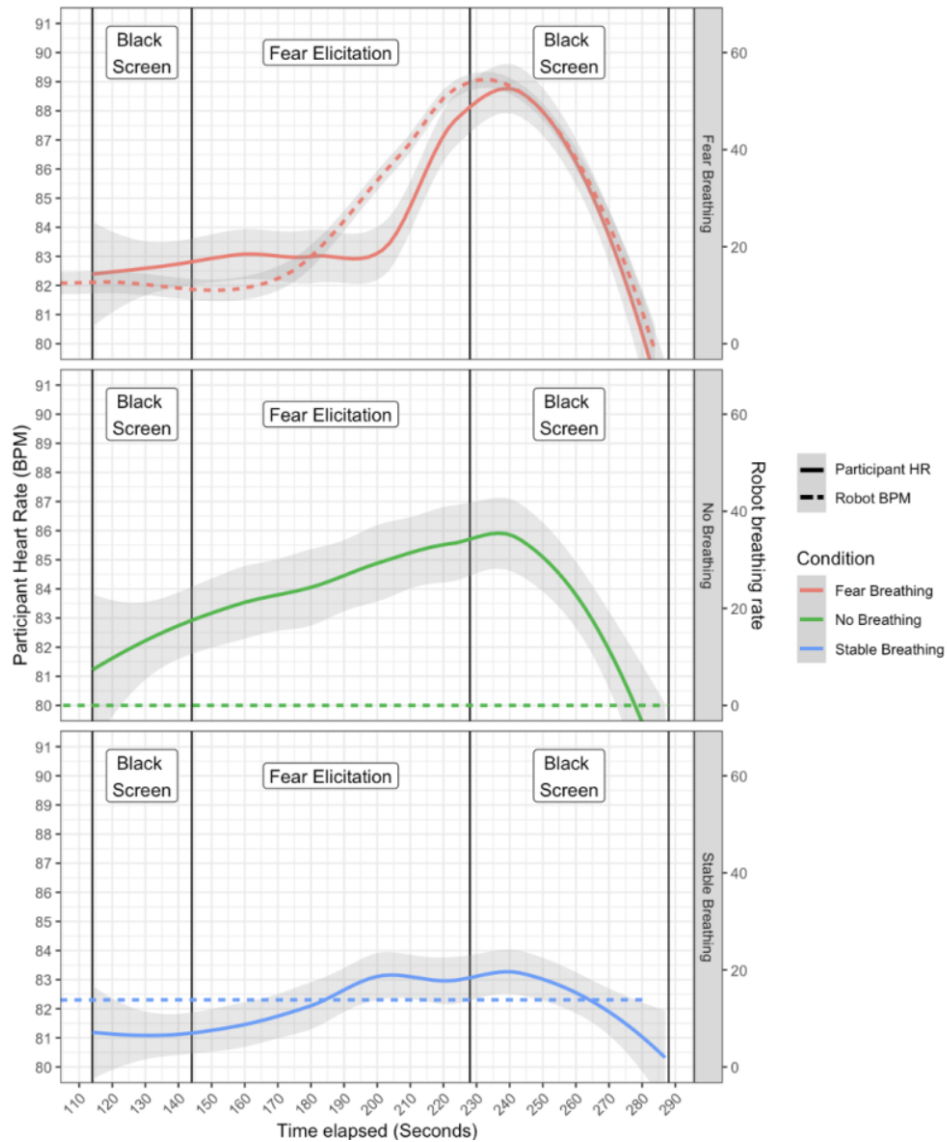


Figure 4.4: Locally Estimated Scatterplot Smoothing (LOESS) lines outlining changes in HR over time (solid line), and manipulated breathing pace of the robot (dashed line) over time, in the Fearful Breathing (top), No Breathing (middle), and Stable Breathing (bottom) conditions. Ribbons indicate 95% Confidence Intervals around local estimates. Note: These data are a combination of data presented in Figures 4.2 and 4.3. The Y-axis on the left corresponds to the participant's HR, whereas the Y-axis on the right corresponds to the robot's breathing rate.

an effect of condition on self-reported positive affect, $F(2,101) = 3.78$, $p = .026$, indicating that participants interacting with the fearful-breathing robot reported significantly lower levels of positive affect than participants interacting with the no-breathing robot, $Beta = .45$, $t(101) = 2.21$, $p = .03$. No difference emerged between fearful-breathing and stable-breathing conditions, $Beta = .16$, $t(101) = 0.78$, $p = .44$, or between the no-breathing and stable-breathing conditions, $Beta = .29$, $t(101) = 1.38$, $p = .17$. These results suggest that, for the most part, any subjectively experienced differences in negative emotion, activation, valence, and negative affect between robot breathing conditions were no longer detectable by the time of our self-reported emotion assessment, approximately five minutes after the conclusion of the robot interaction. However, participants who interacted with the fearful-breathing robot seemed to have experienced minor lingering decreased positive affect. Notably, these results are consistent with those from the HR analyses; by the final moments of the robot interaction (i.e., while viewing the last black screen), all participants had returned to their baseline HR (see Figure 4.3).

4.4 General Discussion

The present research is the first to test whether individuals can recognize apparent fear via touch by detecting expansion and contraction of a robot chest cavity consistent with fear-breathing, and whether this detection influences observers' own fear responses to a fear-eliciting event. Findings demonstrated that, while watching a video clip that reliably elicited fear, individuals who held robots demonstrating a breathing pattern typical of fear perceived the robot as behaving more fearfully, and experienced a pronounced increase in HR, whereas participants who held a non-breathing robot experienced a smaller but still statistically detectable increase in HR, and participants who held a robot exhibiting calm breathing experienced the slightest increase in HR, such that no significant change occurred. These results thus indicate that holding or clinging to others during moments of fear may be ineffective or detrimental for downregulating one's own fear, if those others are also experiencing and expressing fear. Although past research has demonstrated changes in biophysical stress expressions and even pain responses among people in physical contact with robotic or machine-created breathing-like beha-

viours [10, 192, 253, 313], those effects were limited to engagement with robots showing deep and regular movement patterns simulating relaxed or calm human breathing. In the current study, the calm robot breathing condition was designed to emulate that used in prior work, but by also including a fearful-breathing condition we addressed a novel question: how does a robot's accelerated or fearful breathing compare to apparent stable or calm breathing, and to the absence of breathing, in terms of influencing the biophysical stress responses of humans who interact with it? In real-world contexts of humans interacting with other humans or pets (including support animals), it is likely that these interactants would display fearful breathing patterns in response to fear-eliciting stimuli, making the present study representative of how fear contagion is likely to unfold in the real world.

4.4.1 Limitations and Future Directions

It is noteworthy that no differences were observed in self-reported state-level emotion following the procedure. While the absence of greater self-reported fear among participants in the fearful-breathing condition may seem inconsistent with the observed differences in physiology, this is not the case; the absence of self-reported negative emotional effects corresponds to the timing of the observed return to heart rate baseline, because subjectively experienced emotion was measured up to five minutes after the very brief fear-elicitation video. Future research is needed to measure subjectively experienced emotions continuously or intermittently throughout a session like this, to examine whether participants experience greater negative affect or fear in response to fear stimuli when detecting physiological signals of fear through touch. However, adding such an introspective assessment to the current design could introduce new limitations, by interfering with participants' ongoing emotional experience. This addition would also require alternative methods for measuring HR or self-reported emotions, given that participants in the current study had both hands occupied with the robot, and any additional hand movements (e.g., typing or writing) would likely interfere with accurately collecting their HR readings from finger-worn sensors.

Another promising direction for future research is to further compare changes in participants' HR over time for individuals engaging with robots exhibiting stable-

versus still (no-breathing) breathing patterns. In the present study, no significant differences emerged between these conditions when data were analyzed using a multilevel linear model, and including only the first 20 seconds preceding and following the fear-elicitation stimulus. However, this null finding is partly a result of our analysis technique; as depicted in Figure 4.3, which used local estimation (i.e., LOESS lines), HR changes in the stable- versus no-breathing conditions are consistently and significantly different. The failure to capture this difference using MLM is almost certainly due to the fact that our multilevel model takes into account HR data shortly before and after the fear elicitation video, but not during the video, whereas the loess line analysis (see Figure 4.3) takes into account all HR data throughout the procedure. We could not construct a linear model on HR data obtained throughout the entire procedure because these data were severely non-linear, as expected. The results shown in Figure 4.3, in contrast (based on an analysis that included additional data and did not require linearity), are consistent with the suggestion that interacting with a stable-breathing robot while watching a fear stimulus can lower individuals' heart rate, replicating past research [10, 192, 253]. An additional limitation of the present work is that robotic fear-like breathing was simulated using patterns that, though based on past research [28, 220], were dictated by the researchers. Although our manipulation was successful – participants in the fearful-breathing condition perceived the robot to be more afraid – future research is needed to measure the precise breathing patterns demonstrated by humans and pets during fear experiences and manipulate robots to show those patterns. It is possible that additional kinds of fearful-breathing patterns, including breath-holding and gasping, would have similar effects to those uncovered here, but this should be directly tested. Future research should also manipulate and explore the consequences of interacting with robots displaying breathing patterns that are characteristic of a wider variety of emotions, to test which other emotional states are similarly detectable via touch, and similarly contagious when experienced in the appropriate emotionally evocative situations. Future studies would also benefit from examining the effects of interacting with robots that are more human-like than animal-like. Although the use of a robot resembling a pet like a cat or small dog allows us to infer that these results likely apply to individuals using emotional support animals for calming, it remains unclear whether these results apply

to human-to-human contact. Finally, future research should examine the effects of divergent breathing patterns on emotion experience in the absence of an external fear elicitation stimulus. We examined upregulation of emotion during externally evoked fear experiences—an ecologically valid context in which individuals may find themselves touching others displaying accelerating breathing patterns, given that people often cling to others when frightened. However, future work should test whether similar effects emerge when touch occurs devoid of any external emotion context.

4.4.2 Advances in Human-Robot Interaction (HRI)

In addition to advancing knowledge of human emotion experience, contagion, and regulation, these results have important implications for Human Robot Interactions (HRI). Robot-assistance for the downregulation of stress and other negative emotions has many applications and iterations, from the seal PARO for older adults [200, 268] to the Haptic Creature for stress reduction [253]. There also may be noteworthy applications for entertainment or thrill purposes. For example, wearable technologies, virtual reality, interactive movies, and video games might be more evocative and efficacious for eliciting fear when interactants are engaged with dynamic moving machines, such as a controller. Much in the way that vibration can guide emotion experience when playing video games, dynamic expansion and contraction of such machines – such as that used here – might have a similar effect on users’ emotion experience. Future work is also needed to examine whether haptically interactive robots designed with comforting affective touch in mind, like the Huggable (a teddy bear for enhancing pediatric care [280]; the Probo (a plush robot meant for hugging [301]; or wearables like a vest exhibiting squeezing pulsations, might amplify emotion experiences in these contexts. Finally, future research should consider the use of human-robot interactions for clinical purposes, such as flooding or exposure used regularly in cognitive behavioral therapy. The current work thus opens the door to novel methods of enhancing emotional experiences evoked by a range of current entertainment media.

Chapter 5

Machine as Emotion Witness: A Study of Machine Classification of Emotion from Personal Storytelling

Summary

Practical affect recognition needs to be efficient and unobtrusive in interactive contexts. One approach to a robust realtime system is to sense and automatically integrate multiple nonverbal sources. We investigated how users' *touch*, and secondarily *gaze*, perform as affect-encoding modalities during physical interaction with a robot pet, in comparison to more-studied biometric channels.

To elicit authentically experienced emotions, participants recounted two intense memories of opposing polarity in *Stressed-Relaxed* or *Depressed-Excited* conditions. We collected data (N=30) from a touch sensor embedded under robot fur (force magnitude and location), a robot-adjacent gaze tracker (location), and biometric sensors (skin conductance, blood volume pulse, respiration rate).

Cross-validation of Random Forest classifiers achieved best-case accuracy for combined touch-with-gaze approaching that of biometric results: where training and test sets include adjacent temporal windows, subject-dependent prediction

was 94% accurate. In contrast, subject-independent Leave-One-participant-Out predictions resulted in 30% accuracy (chance 25%). Performance was best where participant information was available in both training and test sets. Addressing computational robustness for dynamic, adaptive real-time interactions, we analyzed subsets of our multimodal feature set, varying sample rates and window sizes. We summarize design directions based on these parameters for this touch-based, affective, and hard, realtime robot interaction application.

5.1 Introduction

Social interfaces such as robots, smart cars or game systems must facilitate complex and believable interactions where programmed machines appear to respond to human social cues [88]. Because people often prefer to interact with machines as they do with other people [88], systems may need to understand nonverbal emotional behaviours mediated through naturally affective modalities like touch or gaze. Affective, interactive therapies for anxiety management may use haptically available emotion indicators: touchable robots (baby harp seal Paro [305], teddy-bear-like Huggable [280]) map simple touch gestures to simple emotions. Studies with the Haptic Creature, a zoomorphic robot with an embedded touch sensor array [321], link a large and varied set of touch gestures to nuanced emotion expression.

Machine recognition of human emotion presents methodological challenges surrounding measurement instruments, study task framing, and computationally modeling emotions [38]. Training data behavior should reflect that of an interaction “in the wild”, i.e., spontaneous emotion [93]. The emotion model should accurately describe that person’s state. Furthermore, while people can be differentiated by idiosyncrasies in their touch behaviors (a *touch signature* [41, 87]), this also makes it difficult to generalize the connection between emotions and associated touch behaviors: the extent to which individuals exhibit similar touch behaviours during similarly labeled emotional states is unclear.

Here, we wish to enable machine recognition of human emotions for touch-centric social robots, with therapeutic applications in mind. Touch interactions can affect emotional state: the Haptic Creature’s motion lowered *anxiety* in users

who were stroking it on their laps [252], based on biometric indicators. This suggests physiological benefits analogous to those conferred by animal-assisted therapy [14, 15, 208, 230] – especially valuable where patients are unable to engage with real animals. However, this requires unobtrusive sensing, e.g., through already-occurring touch.

Gaze is another unobtrusive modality that could improve recognition performance. Since the points where a user’s gaze focuses on a computer display can indicate feelings of curiosity or boredom [139], we posit that gaze as an indicator of visual attention could help determine when a user is focusing on the robot pet and thereby predict affect. Specifically, we compare the combination of touch and gaze to key biometric channels which have been well-researched in association with various emotions [155, 161].

To investigate these ideas, we set touch as the primary interaction modality in order to leverage the natural human inclination to express emotional closeness with physical contact. Gaze has also been shown to capture emotion data [139], and both (touch and gaze data) can be collected without the disruption of physiological sensors. Previous work has shown that affect-related information can be extracted from emotionally-directed touch gestures such as *Excited*-stroking and *Depressed*-rubbing [5]. However, identifying a gesture as ‘stroke’ vs. ‘rub’ is insufficient for revealing the user’s emotional state while performing that gesture [5]. Furthermore, these studies collected “intent” data, where the emotions were *acted out to* a sensed robot, but not necessarily *experienced by* a participant. We needed a model built from data of participants who are truly experiencing the emotions being studied.

5.1.1 Approach and Research Questions

The central purpose of this paper is to narrow the design space of an emotionally interactive robot pet’s computational system for predicting an interacting user’s emotion: touch-supportive sensing modalities that balance accuracy with ease-of-use; a training procedure that generates truly felt emotional sample data; and an appropriate classification model for touch behaviour in a computationally restricted environment.

To elicit naturally felt, spontaneous human emotion (hard to do in a lab set-

ting [93]), we asked participants to interact with a robot while they relived a significant emotional event, touching it without constraint during the task. This approach departs from previous work [5, 321] that attempts to direct touch behaviours and gestures, i.e., by asking a participant to *pat* the robot *as if* they were *scared*. Relived emotion or emotion recall is regarded as a way to elicit true experiences of emotion [80, 179].

We are interested in touch and gaze as modalities that support low-cost, low-intrusion sensing apparatus and explore their viability in comparison to biometric data. To that end, we compared affect measures derived from touch interaction with a robot pet with the more studied but intrusive reference point of biometric indicators, and investigated how recognition performance can be improved with gaze data. Furthermore, analysis methods that originate from social touch gesture classification are well documented [5, 87, 150]. We calculate features from force magnitude and touch location [41, 87, 150] as well as frequency [5] (referred to herein as pressure-location domain and frequency domain respectively) for emotion classification in *touch*. To minimize overlap in label interpretation, we collected and evaluated machine recognition of four emotions (*stressed*, *excited*, *depressed*, and *relaxed*) – quadrant extrema of Russell’s dimensional affect model [237].

Choice of the Random Forest algorithm (RF) is motivated by our need for a classification system that performs well with social touch behaviour [5, 41, 87, 96, 152, 288] for our interactive robot pet application. We want to explore the feasibility of realtime emotion prediction from touch interaction with a emotionally interactive robot pet, where we anticipate being compute-restricted. Thus, we chose a computationally simple model favouring flexibility to accommodate quick training and customizable rebuilding.

We specified four main research questions for this study.

RQ1 Modality Effectiveness: How does touch or touch + gaze compare with biometrics in classifying affect? What minimal feature set optimizes performance accuracy?

Touch can be a natural avenue for communicating affect, but to use it computationally, we must access the encoded emotions and consider the relative performance of touch alone and with multimodal support. Gaze, also known to encode affect-

ive content [139], could supplement emotional signals from touch. Multimodal datasets are likely to provide a more complete picture than touch alone, due to asynchronous activation, or interaction information.

We expect *classification accuracy to improve with increased modality support*. We thus ask whether the combination of touch and gaze is a viable substitute for the more intrusive sensing apparatus required of tracking biometric signals.

However, multimodality increases compute time and phase delays, potentially undermining real-time feasibility. To optimize tradeoffs, we analyze each feature in terms of repeated occurrence in automatically-selected best-feature subsets. Finally, we suggest an optimal touch-with-gaze feature set, assessing both the *pressure-location domain* and *frequency domain*, hypothesizing that *classification accuracy is best where features are present from both domains*.

RQ2 Individuality: How important is system calibration and knowledge of user in affect classification?

Social touch gesture studies suggest that because individuals have distinctive ways of physical, expressive interaction with objects, recognizing *identity* is realistic [41, 87]. Thus a system that has learned a specific user’s behaviour may be better at gesture recognition. Leveraging this result for affect, we assess how well the system can distinguish Participant – high performance suggests high individuality – then perform Emotion classification across three different levels of system knowledge of participant (hereby referred to as *participant knowledge*) and discuss results. We expect that *recognition rates will increase with greater participant knowledge*, i.e., participant-labelled data where instances from the same individual are in both training and test sets will yield the highest classification accuracy (subject labels used as a feature in subject-dependent classification); and lowest accuracy will coincide with testing and training on different individuals (subject-independent classification).

RQ3 Sample Density and Realtime Responsiveness: Is classification during continuous sampling robust to interruptions in signal, and to sample size variation?

Outside of polling rate, we define *sample density* across two window dimensions: (1) size and (2) adjacency. We investigate the accuracy trade-offs of various *win-*

down sizes – which represent the time intervals of continuously sampled data. In the context of an interactive robot, longer windows gives the system time to respond, employs less computation resources and allows for the capture of ”slow” behaviours. But where the window is too long, we introduce inappropriate response delays. For example, if our robot body is struck, it needs to present a behaviour demonstrating an immediate reaction. While shorter windows may help with the agility needed for interactive scenarios, the higher throughput requires more computational resources and may not recognize the slower developing interactions.

Window adjacency refers to continuity of time series classification data. Since adjacent windows share more characteristics than distant samples (temporal dependence), we ask about the effect of non-continuous or ‘gapped’ data collection under weak or interrupted signal conditions. Removing adjacent instances allows us to quantify any effect from a dropped or intermittent signal as well as the likelihood of overfitting due to recency-based similarities, particularly when using easy-to-build classification models (like Random Forest) without parameter tuning. Here, we leave time-series analysis for future work and focus on the influence of sample density on accuracy. In order to construct early specifications for a touch-cognizant robot, we explore the trade-off between computational load and classification robustness.

We examine the influence of window size and continuity by aggregating data instances in four window sizes and comparing classification accuracy of the same data set. We downsampled *with “gap”* by dropping 2s of data between windows so adjacent windows are not evaluated) and *without gap* data (adjacent windows are included in the training and test sets). We posit that across both parameters, *reducing sample density reduces classification accuracy*, anticipating the worst performance for small windows with gapped data.

RQ4 Experimental Paradigm: How well does our protocol corroborate existing relived emotion techniques to elicit genuine emotion in a controlled laboratory setting?

For affective communicative systems to work under real conditions, they must be trained on data from authentic and spontaneous emotion. Consistently producing *truly experienced* emotions in an artificial setting (and valid training data) is a

fundamental challenge in emotion research [57].

We develop a means of implementing a touch variant of relived emotion techniques described in [57, 179, 180] and use self-report measures to explore how our experimental controls influence the *authenticity and intensity of the experienced emotion* generated within a controlled set-up.

5.1.2 Contributions

Through our research questions, we examine the design space of an affect classification system for an emotionally-interactive touch-centric robot. Specifically, we contribute:

1. *A comparison of affect classification performance* of touch data, with and without gaze support, to biometrics in *experienced-emotion* interactions; and a recommendation of data features from frequency and traditional pressure-location domains in emotion classification.
2. An assessment of subject-independent vs. dependent classification; and a *proposal for building a custom personalized system* at various levels of participant knowledge.
3. *An analysis of data factors* to balance classification robustness with computational effort and phase delay, for real-time applications.
4. Through demonstration and evaluation of an ecologically valid elicitation technique (emotional recall) for studies on machine touch recognition, we *assess the methods, models, and task framing required to increase confidence in generating true experienced emotion in a lab setting*.

In the following, we survey previous work, motivating our emotion elicitation method and contextualizing affect classification from each of touch, gaze, and biometrics; then describe our experiment and analysis. We report results that span all our data experiments to target the influence of: multimodal data vs. touch alone, participant knowledge, sample density, feature set; and assess emotional experience from participant reports. We discuss our findings and ground them in implications for relevant applications.

5.2 Related Work

5.2.1 Targeted Emotion Set

Russell’s circumplex model plots affect on arousal (activation) and valence (pleasantness) axes [239]. While valuable in its conciseness, the dimensional model requires we assume (1) emotion labels will be interpreted consistently by every participant at any time; and (2) the axes are truly orthogonal.

Consider the emotional context of approaching the axes or origin when working with such a model: the state of (0,0), presumably a state of full neutrality, may not be meaningful. For example, independent movement, i.e., directly along axes, implies increasing an emotion arousal without changing valence, which belies personal experience. As such, many [63, 110, 309] opt to discretize the 2D space into a grid and rotate it by 45° , such that experimental materials and tasks are aligned with the diagonal axes, namely (high arousal, high valence) \leftrightarrow (low arousal, low valence) and (high arousal, low valence) \leftrightarrow (low arousal, high valence).

Relevant published studies are not consistent in emotion labels chosen to cover the affective space, making comparison between studies of even common modalities problematic. Understandably, papers utilizing information of gaze use attention-related emotion sets – e.g., *Anxiety*, *Boredom*, *Confusion*, *Curiosity*, *Excitement*, *Focus*, *Frustration* [240]; papers utilizing touch try to span the human experience, namely *Anger*, *Fear*, *Happiness*, *Sadness*, *Disgust*, *Surprise*, *Embarrassment*, *Envy*, *Pride* [114]. Yet another method is to partition Russell’s affect grid as discrete labels: touch emotion recognition has previously used nine labels¹, while biometric recognition has used four labels corresponding to the quadrants of Russell’s grid: *Stressed*, *Excited*, *Depressed*, *Relaxed* [155]. We have elected to use the same four named emotions for consistency with other biometric classification studies, enabling comparison with touch and gaze.

5.2.2 Elicitation of True Emotion

Our motivating applications center on a social robot that must react to authentic human emotions as they occur in lived experience. In the lab, one unsatisfying approach is to ask participants to imagine and simulate a reaction: (“*Imagine feeling anger, then express it to our robot*”). For example, to collect the data used

¹Emotions for classification by touch differentiates emotions in the quadrant borders, namely: *Distressed*, *Aroused*, *Excited*, *Miserable*, *Neutral*, *Pleased*, *Depressed*, *Sleepy*, *Relaxed* [5].

in [321] and [5], participants were presented with a list of emotions that they acted out by touching a robot, but this does not equate to experiencing it. The difference between expressions of acted and experienced emotions can be significant and counter-intuitive: e.g., truly experienced frustration is often accompanied by a smile, but this is rarely the case for acted frustration [127].

Experienced-emotion studies are difficult to construct. Entertainment media, e.g., emotionally evocative music and/or video, has been employed in emotion elicitation [155]; however, it can be difficult to validate stimulus media. Following the approach of [80, 179] who found that relived or recalled emotion generated genuine spontaneous reactions, we prompted participants with an emotion word and asked them to recount the story of an intense experience with modifications described in Methods.

5.2.3 Recognition Modalities

Touch: We can measure touch as force magnitude (pressure) and location – dimensions used for gesture recognition as well as for control directives using trackpads and touch screens. Social touch gesture studies report prediction accuracies ranging from 53% (chance 7%) [150] to 86% (chance 11%) [87] depending on collection and classification methods (Bayesian classifiers in the former and random forest in the latter case), and like affect studies in general, have no consistent standard. Still, these prediction rates on defined gestural subsets suggest that social touch may be used as directives in systems with embedded recognition systems.

Accurate *emotion* recognition is more difficult. Human recognition of human emotion through touch reaches 59% accuracy (chance 8%) [114]. Machine classification has demonstrated 36~48% accuracy (chance 11%) [5] depending on inclusion of participant information. Both studies utilized emotion *intent*, not *experience*.

Gaze: Our eyes give affect cues discernible with eye tracking technology, making gaze behaviour an easily accessible emotion-embedding modality to pair with touch without hindering interaction. Like touch, gaze detection technology collects eye behaviour at the focal location and does not require participants to wear sensors on their body. [217] studied the effect of emotional auditory stimulation on

pupil size variations, finding that negative and positive stimulation resulted in larger pupil dilation than neutral stimulation but did not differentiate stimulus valence. Other factors, such as changes in luminance [117], can also affect pupil dilation.

An alternative is to analyze where a person is looking. [139] tracked students' gaze when they interacted with a graphical intelligent tutoring system; fixation and saccade features revealed that curious and bored students looked at different interface areas – e.g., engaged students looked more at the table of contents. Overall, boredom and curiosity could be predicted with 69% and 73% accuracy respectively.

We could not find literature on the use of human gaze *point* in classifying emotions using the valence/arousal model. Gaze point is related to boredom and curiosity, and low arousal is correlated with decreased saccadic velocity [74], but can gaze express arousal change too? Does gaze point move more during excitement? Compared to pupil size variation measurements, gaze point can be measured in a less controlled environment (lighting and luminance impact data quality less) with relatively inexpensive tracking technology. Thus, we utilize the Cartesian coordinates of user gaze point in our own classification analyses.

Biometrics: Blood volume pulse (BVP), skin conductivity (SC) and respiratory rate (RR) have been widely used to confirm emotion detection in other modalities – facial expressions [161], affective audio [155, 202], gaze behaviours [120], and touch behaviours [252]. Heart rate variability has been utilized in emotion classification [9, 145, 252].

Like others, we employed three basic signals (BVP, SC, RR) to calculate a set of derived features based on heart rate variability (HRV), breathing rate variability (BRV), or both, such as heart beats per breath. This data is most appropriately compared with studies where emotion elicitation is based on true experience and uses the same emotion sets. For example, [155] uses validated music excerpts to generate authentic responses crossing four musical emotions (positive/high arousal, negative/high arousal, negative/low arousal, positive/low arousal), and reports affect recognition rates between 70% and 95% (chance 25%), with higher rates when participant knowledge is included.

5.3 Methods

We asked participants to recall emotionally intense experiences, while interacting with our static (non-mobile, unmoving) robot pet as a tangible focus for emotional interaction. We collected touch, gaze and biometric data; and emotion self-reports before and after each emotion. Of 30 campus-recruited participants (mean age 25.4 years, $\sigma=5.4$), 14 identified as female and 18 had corrected vision. Participants were compensated \$20 for a ~ 60 minute session.

In the following we detail data collection setup and procedure, and describe data pre-processing, feature extraction, and analysis of the study’s independent parameters (*window size*, *inter-window gaps* and *participant knowledge*).

5.3.1 Data Collection

To facilitate emotion elicitation during memory recall, we prioritized participants’ comfort. We placed the gaze tracking system coincident with touch site, since the robot body is the focal site for both modalities.

Configuration and Room: We conducted the experiment in a sparsely furnished medium-sized office with a window with a pleasant view. Participants sat, back to the door, comfortably in a half-prone position on a couch, for comfort and to reduce large-scale body movements (Figure 5.1). An experimenter was in view of the participant except during emotionally intense parts of the session, as described below.

We placed the gaze tracker (designed for mounting beneath a computer monitor) below an angled, monitor-sized board on which we placed the robot, all in comfortable reach of the participant. We fixed the robot position to prevent it from being picked up or substantively moved around to avoid interference with gaze tracking (Figure 5.1). By coincidence, all participants were right-handed (though the set up was designed to accommodate both right- and left-hand dominance) and we omit a discussion on handedness.

Touch Sensor on a Passive Robot: Figure 5.1(a) shows the robot’s and sensor’s construction. We used a custom flexible touch sensing apparatus previously described in [41], which has been validated as for the ability to capture social touch gestures. Similarly to [87, 151], it can detect 5g \sim 1kg of weight with resolu-

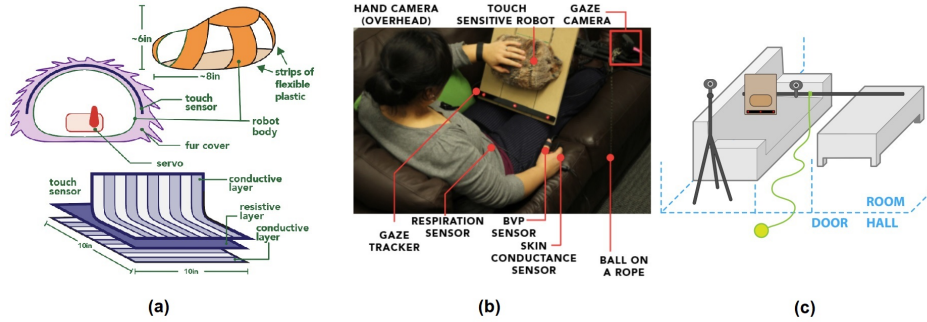


Figure 5.1: Study setup overview: robot description and participant experience. **(a)** The robot was constructed from pliant plastic sheets actuated by a pulley, covered with a custom touch sensor, then jacketed in furry fabric to invite touch [37]. It was stationary during the study to eliminate reaction to robot motion. **(b)** A participant sits supported by pillows and facing the gaze tracker, one hand on the sensor-clad, stationary robot, biometric sensors on chest (RR), thumb (BVP), and index / ring fingers (SC) of resting hand. **(c)** A schematic of the study room, depicting camera locations relative to where the participant sits by the robot platform.

tion of 10×10 inches at one taxel per square inch². As with [5, 87], we specified fingerpad-size taxels (touch pixels): emotion tasks in touch generally incite broad rather than precise movements [114]. While higher resolution sensors are needed for precision tasks (e.g., for touch screens, trackpads, or teleoperative mimicry [273]), here we are concerned with cost, sensor flexibility and computational efficiency.

Forming a 10-by-10 grid, this fabric-based device can sense multiple simultaneous touches (multitouch), registering varying pressures on each taxel scaled to 1024 levels and polling at 54Hz. This resulted in 54 frames of 100 cells per second, each reading a touch pressure value in [0-1023].

The “bot” was assembled in layers. The interior was a compliant structure of flexible binder plastic, roughly the size and weight of a football. The robot’s body and passive feel were designed to invite touch as an ambiguous mammalian form

²Built from commercially available piezoresistive and conductive fabric. Fabric is commercially available at www.eeonyx.com.

that does not resemble any definitive animal in order to remove behaviour expectation [35, 37]. Movement was disabled here to reduce confounds from novelty effects, sounds, or expectations. The touch sensor was wrapped over the structure, affixed with velcro. Finally, the sensor was covered with a uniformly-textured short, soft brown minky fabric (such as that used in baby blankets; described as “*pleasant to touch...[and] reminded me of my chocolate lab’s head*” – P04). To minimize visual clutter, all sensors were wired through the robot platform and gathered in a compact tether for connection to a single laptop.

Gaze and Biometric Sensors: We sampled gaze behaviour via a Tobii EyeX gaze tracker at 60Hz – as with our touch data sampling (Figure 5.1). We gave no specific instructions regarding gaze direction, but informed participants that gaze data collection worked best when they were facing forward and did not make large body movements.

We collected three biometric signals using the pre-packaged Bio-Graph Infinity Physiology Suite³, namely blood volume pulse (BVP), skin conductivity (SC), and respiratory rate (RR), all at 2048Hz. Following established procedures [155], these were expanded to include features on heart rate variability (HRV), breathing rate variability (BRV), and cross-signal indicators such as heart beats per breath.

Participants wore a respiration band around their chest, with the closest fit that did not impede breathing. Once the participant was comfortably seated, we positioned the BVP sensor at the thumbpad, then positioned the SC sensors on the index and ring finger pads. Both BVP and SC sensors were held in place by a small velcro band on the right hand (not used for touching the robot).

Video Data: We video-recorded participants’ hands and face to supplement missing gaze or touch data. For participant privacy, no sound was recorded. The hand camera was placed behind, and the face camera on the right of the participant. Figure 5.1 shows placement of the gaze tracker.

Emotion Labels: Genuine emotion is taxing. To minimize fatigue, we administered just two emotions per participant, based on discussions with field experts, piloting and literature. The second emotion task was determined by the first; par-

³System manufactured by Thought Technology Ltd. FlexComp ∞ SA7550 Hardware Manual can be found through manufacturer website at <http://bit.ly/29A5NIC>.

Table 5.1: Experimental procedure and data acquisition.

Step	Description (duration)	Data or Output
(1) Intro	Describe study tasks	informed consent
	calibrate sensors	verify data quality
(2) Neutral 1	Read neutral text (5 min)	biometrics
	Self-report	emotional state
(3) Emotion 1	Calibrate gaze/touch sensor	calibration logs
	Recall memory ($\mu = 4.23$ min, $\sigma = 3.09$)	biometrics, gaze, touch
(4) Neutral 2	Self-report	emotional state + authenticity rating
	Read neutral text (5 min)	biometrics
(5) Emotion 2	Self-report	emotional state
	Calibrate gaze/touch sensor	calibration logs
(6) Debrief & Interview	Recall memory ($\mu = 4.23$ min, $\sigma = 3.09$)	biometrics, gaze, touch
	Self-report	emotional state + authenticity rating
(6) Debrief & Interview	Interview	qualitative data
	Self-report	emotional state

participants experienced either *Stressed - Relaxed* OR *Depressed - Excited*, counterbalanced. The four named emotions [*Stressed, Relaxed, Depressed, Excited*] comprised the emotion label set and validated via self-report on intensity and authenticity and coordinates on Russell’s affect grid [237].

Procedure: Table 5.1 summarizes our study procedure, in which neutral steps delineated experiment steps. Emotion tasks were counterbalanced across participants.

Introduction and Calibration: To reduce novelty effects, we introduced the robot, invited touch exploration, described the robot including its sensing abilities, and explained that its movement was disabled. We then calibrated all sensors.

Neutralization and Self-report: For each stage, we first presented an emotionally neutralizing reading task, wherein the participant read aloud from a short report from a technology magazine for ~ 5 minutes. We instructed the participant to read each word, told them that no questions would be asked of the readings, and encouraged them to let go of residual emotions from their day.

We then asked the participant to report their current emotional state. Before each emotion self-report, an experimenter explained or reminded the participant of concepts of arousal and valence, answered questions about reporting emotional

state, and showed them how to indicate their current emotional state on a form displaying Russell’s [239] 2D affect grid varying in arousal and valence [57]. This self-report was repeated before and after each neutralizing and emotion task. For emotion tasks, participants were also asked to rate how strongly or authentically they experienced the emotion, compared to the original incident.

Reliving Emotion Task: We next asked the participant to recall an emotionally intense memory pertaining to an assigned emotion word {*Stressed*, *Excited*, *Relaxed*, or *Depressed*} as they interacted with the robot. To elicit strongly emotion-influenced touching, we invited them to relive the emotion as intensely as possible while keeping their non-instrumented hand on the robot. We explained that audio recording was disabled in the video camera and we could not hear them speak from outside the room. They received no other touch instruction or reminder. After we left the room, they described their memory with its associated feelings to the robot in any language, at a volume of their choosing. The participant indicated task completion by pulling a signal rope. Data was collected for a single recalled memory (duration $\mu=4.23$ min, $\sigma=3.09$ min).

When the rope was pulled, the experimenter returned and administered the self-report grid, then repeated the steps for the second set of neutralization and emotion tasks.

Debrief and Interview: We conducted a short debriefing interview to learn of any unexpected eventuality during their experience, and ensure that participants were comfortable, emotionally stable, and departing in an emotional state no worse than when they arrived. We provided university counselling contacts after we found in piloting that participants could become distraught during this protocol.

5.3.2 Features, Pre-Processing, Extraction & Analysis

We recorded touch, gaze, and biometric data for affect classification features (see Table 5.2 for a full list). Here, we describe the feature extraction process.

Distribution statistics: We included conventional touch statistics [5, 41, 87]: min, max, mean, median, variance, total variance, area under the curve (AUC) for location X- and Y-centroid and touch pressure. Touch pressure is computed by frame:

Table 5.2: Summary of features extracted from *touch*, *gaze*, and select *biometric* signals.

FEATURE	SIGNAL	#
TOUCH (54Hz)		
<i>distribution</i> : max, min, mean, var, total var, AUC (Area Under Curve)	Xcentroid, Ycentroid, frame pressure	21
<i>frequency</i> : peak count, fundamental frequency, amplitude max, mean, var & total var	Xcentroid, Ycentroid, frame pressure, pressure of centroid cell + 8 nearest neighbours (9 vals)	72
GAZE (60Hz)		
<i>distribution</i> : max, min, mean, var, total var, AUC	X, Y, saccade length, velocity, fixation duration	25
<i>sample counts</i>	total samples, on/off-robot, off-on robot ratio, rate within platform range, saccade count, saccade rate, fixation count, fixation-saccade ratio	9
<i>frequency</i> : peak count, fundamental frequency, amplitude max, mean, var & total var	X, Y	12
BIOMETRICS (2048Hz)		
<i>summary statistics</i> : mean, median, variance	<i>Blood Volume Pulse (BVP)</i> : amplitude, high frequency power (FP), low FP, very low FP, heart rate, inter-beat interval, peak amplitude	228
	<i>Skin Conductance (SC)</i> : mean, epoch mean	228
	<i>Respiration pattern</i> : abdominal amplitude, respiratory rate, period	228
Thought Technology's commercially available calculations were used for biometric feature extraction: http://www.thoughttechnology.com		

pressure values per capture of the 10x10 sensor. For the centroid, we found the cell containing the coordinates of the touch-pressure centre of mass (X-centroid, Y-centroid); i.e., the weighted average of all taxels in a frame based on their row and column locations, or (X, Y) coordinates respectively. Gaze focal location (x,y) and biometric channels of blood volume pulse (heart rate), skin conductance, and respiration rate were similarly calculated.

Frequency statistics: Based on prior indications of promise [5], we extracted frequency-domain features to assess how well they encode emotion content. We calculated six frequency statistics for 12 touch signals and the same six for two gaze signals. We directly calculated frequency-domain touch and gaze features, and used Thought Technology's pre-packaged signals⁴ for biometrics.

Feature Extraction

We calculated distribution and frequency statistics for touch and gaze. For biometric features, we relied on prepackaged calculations but also computed simple statistics (mean, median, variance) for insight into distribution characteristics. Table 5.2 summarizes the full feature set.

Touch features: We reprised known procedures for social touch recognition by constructing three parameters [41, 87]: *touch pressure* (sum of pressure readings from taxels in frame); and *column* and *row centroids* (weighted measure of row, column centres of mass based on frame taxel pressure, or X-centroid and Y-centroid respectively). We computed 7 statistics per pressure parameter, for 21 features.

For frequency-based features of emotive touch, we performed a Fast Fourier Transform (FFT) of the three frame-level pressure and the centroid coordinates (x,y) described above; and then calculated 6 frequency statistics for each as well as the pressure readings from the centroid cell and its eight nearest neighbors [5], comprising 72 more touch features in the frequency-domain.

Gaze features: From the gaze data, we collected raw (X, Y)-coordinates of focal points from the Tobii eye tracker and calculated 34 features: distribution statistics for each of {focal coordinate pair (X-, Y-location), saccade length, velocity, fixation duration} as well as 9 summary features of gaze presence and location including saccade and fixation ratios. We used Salvucci’s I-VT algorithm [245] to differentiate between fixations and saccades. Gaze samples with point-to-point velocities $<30^\circ/\text{s}$ were classified as fixations and those with velocities $\geq 30^\circ/\text{s}$ as saccades. We calculated 6 frequency statistics for gaze data on the 2D focal location, generating 12 frequency-domain gaze features.

Biometric features: We computed mean, median, and variance across all signals provided from the Thought Technology physiology suite, including both base signals (BVP, SC, RR), and channels dependent on the original signals (HR, HRV, IBI, etc.), for a total of 228 features across 76 channels.

Data Instances / Partitioning on Independent Factors

Each data instance is comprised of a list of touch, gaze, and biometric features computed across a single time window. We omitted windows that provided insufficient samples for FFT (<10) for any modality⁵ – generally due to gaze data loss when gaze was outside of the tracked area. We partitioned our data and analyzed how key computational factors influence classification accuracy: window size (data density), inter-window gaps (continuity), and participant knowledge (content) (Table 5.4).

Table 5.3: Data instance count by Emotion and Participant.

Pno	Depressed	Excited	Pno	Relaxed	Stressed
P01	5426	3114	P02	1323	1309
P05	2251	2559	P04	2308	2656
P07	1755	1842	P06	3292	1895
P09	1726	7694	P08	2415	3888
P11	3166	3217	P10	884	1667
P11	2574	1557	P14	2421	2239
P15	3492	4275	P16	1880	2030
P17	1428	1337	P18	1479	1806
P19	1322	1286	P21	922	1290
P20	1608	1286	P23	1668	3235
P22	2070	1824	P25	2567	1735
P24	5873	10232	P27	954	2084
P26	5050	3722	P29	2392	2119
P28	1268	960	P30	1268	1783
P31	755	563	P03	<i>data corrupted</i>	
P32	1704	2148	P12	<i>data corrupted</i>	
μ (σ)	2591.8 (1584.7)	2976 (2594.9)	μ (σ)	1840.9 (740.4)	2124 (714.8)

Participants ($N = 30$) were allotted as much time as needed for their emotion task – telling or reliving an intensely emotional story. Because timing was not restricted, the data instance count for each emotion word varies.

⁵On average, usable data instances dropped by 36% with shorter data windows being more affected.

Window Size: Impact of window size on classification is crucial for compute-constrained real-time gesture classification. 2s windows (54Hz, or 108 frames) have been used to capture touch gestures [87]; however, human hands and fingers can move at $\sim 100\text{-}200\text{ms}$ [212, 269].

We therefore partitioned data in 2s non-overlapping windows and extracted features for training and test instances. Each data instance has features extracted from a 2s window to build a classification model. This partitioning and feature calculation were performed on the same data at other window lengths, resulting in four distinct sets of data instances at [0.2s, 0.5s, 1s, and 2s] windows.

Inter-Window Gaps: Even though our Random Forest classification model treats instances without temporal dependence, we consider that temporally-neighbouring instances can be exceedingly similar, particularly in the smallest windows. We investigate whether, and by how much, recency effects influence accuracy rates by adding 2s gaps between instances thereby eliminating adjacent instances. We compare classification performance of the data with and without this artificial gapping (gapped vs un-gapped data.)

Participant Knowledge: We report accuracy for **emotion** classification across three levels of the classifier’s knowledge of the participant in increasing information order:

1. **No participant knowledge** – *subject-independent* classification simulates the task where an interactive system’s emotion model cannot be trained on all possible users. E.g., a robot in a museum or institutional context must be modelled on a training set that could not include all possible users, who are not known ahead of time.
2. **Implicit participant knowledge** – this *subject dependent* system simulates a classification task where the interactive system’s emotion model has been trained on all expected users before classification but not explicitly informed which data is associated with the current user. We imagine a system that lives in a limited private domain, where all users have completed a calibration period, informing the model’s training set.
3. **Explicit participant knowledge** – the training set includes participant labels

Table 5.4: A motivating overview of analysis factors.

WINDOW SIZE: [0.2s, 0.5s, 1s, 2s]	
Description	Data was all sampled to 54Hz. Window size is the length of time over which a feature is calculated. e.g., a two-second window has 108 samples.
Implication	With a static sample speed, shorter windows simulate a system with faster update cycles, resulting in less information per window, but faster system response.
Question	How do accuracy rates change with different sample sizes?
INTER-WINDOW GAPS: [Without gaps, With 2s gaps]	
Description	With no gap, all windows are calculated contiguously, i.e., every window is directly adjacent to the one previous. With gap, after every window is calculated, two seconds of data is discarded.
Implication	Social touch gestures take a little under a second to make [87] so a 2s gap increases the likelihood that each window captures different gestures.
Question	How robust is the system to data loss?
PARTICIPANT KNOWLEDGE: [Explicit, Implicit, None]	
Description	The system may select participant labels if included in the training data. We have three levels of participant knowledge: participant labels included, participant labels excluded (both subject dependent), all participant data excluded (subject independent).
Implication	When labelled, the system can tell whose emotions it is attempting to predict. When unlabelled, the system still has knowledge of the participant’s behaviour, but cannot determine from whom. The most challenging case: testing on a participant’s data without her training samples.
Question	How much does <i>a priori</i> identification of an individual influence classification accuracy?

as a feature (subject-dependent where instances are attributable by subject). This system knows whose emotions it is attempting to classify and loads a personalized emotion model for each user.

We also ran **participant** classification to determine not only how well these feature sets can determine *what interaction* was performed, but also *who* performed it.

Classification

Here we summarize the classification tasks: predicting *emotion* and *person* experiencing the emotion while experimenting with data instances comprised of our statistical features and varying window size, inter-window gaps, and participant knowledge. For literature comparison, we report classification accuracy as the ratio of correctly classified instances over all instances as well as multi-class weighted

F1-scores based on the instance count of each class.

We used Weka, an open-source machine learning platform [108], for k -fold cross-validation (CV) using a Random Forest (RF) classifier – so chosen for its known efficacy for touch recognition [87, 152] and low training and computational threshold – to assess classification accuracy on both pressure-location and frequency domain features. We chose a relatively moderate value of $k = 20$ for our CV, to support comparison with other studies which have shown this method to be effective in touch classification [5, 41, 87, 150]. We included subject-dependent tests for models trained on all participants as there is no restriction on whose data instances are included as training or test data, so long as the same data instance is not in both.

Subject-independent Emotion Classification: For subject-independent analysis (no participant knowledge), we use two types of Leave-One-participant-Out (LOpO) classification: (1) one participant’s data is left out and the training set includes *all other participants (LOpO-All)* (i.e., training $N = 30 - 1$) and (2) one participant’s data is left out and the training set includes all other participants *who performed the same emotion tasks (LOpO-Half)* (i.e., training $N \approx 15 - 1^6$).

LOpO-ALL simulates a system that has no knowledge of a new user and has been trained on all emotional touch behaviours (chance $\approx 25\%$). **LOpO-HALF** simulates a system that has no knowledge of a new user and has been trained only on the 50% subset of behaviours this user *will* be performing (chance $\approx 50\%$).

Subject-dependent Classification: Given the highly individual nature of the touch behaviours we observed, it is possible to expect LOpO classification to perform at or near chance. We also performed CV for conditions classifying:

1. *Participant:* represents a system trying to identify *who* is performing the interaction.
2. *Emotions given explicit participant knowledge:* participant labels are included as a feature;
3. *Emotion given implicit participant knowledge:* participant labels are omitted.

⁶Test sets are comprised of data instances from participants who performed *Stressed* and *Relaxed* ($N_{SR} = 16$) or ones who performed *Excited* and *Depressed* ($N_{ED} = 14$)

Table 5.5: Weighted F1-scores from 20-fold cross validation varying factors of Gap(+/-), Participant Labels(+/-), and Window Sizes (0.2s, 0.5s, 1s, 2s) on touch *T*, gaze *G*, and biometric *B* features, classifying emotion ($25\% \leq \text{chance} < 50\%$). Classification accuracy is within 0.003 from these values. Weighted F1-scores that are from 0.01 to 0.03 below classification accuracy are indicated with *.

	Win	Participant Labels-				Participant Labels+			
		T	G	TG	B	T	G	TG	B
Gap +	0.2s	.666	.412*	.704	.997	.871	.744	.884	1
	0.5s	.693	.448*	.735	.996	.881	.773	.897	1
	1s	.719	.489*	.759	.996	.886	.781	.909	.999
	2s	.566	.465*	.597	.892	.793	.651	.765	.942
Gap-	0.2s	.754	.475*	.822	1	.923	.788	.944	1
	0.5s	.761	.505*	.823	1	.921	.803	.939	1
	1s	.768	.530*	.821	1	.921	.811	.937	1
	2s	.761	.569	.815	.999	.918	.813	.931	1

5.4 Results

Consistently with past studies on biometric-based emotion classification [155], our biometric data alone gave accuracy rates from $\sim 90\%$ to near 100% (Table 5.5).

This section describes our results from running classification using our full feature set on emotional touch and gaze behaviour across a number of experimental conditions (compared to that of biometrics alone) (Table 5.5). We also look at subject-independent tests of emotion classification which also employed the maximal combination of modalities (touch + gaze + biometrics) (Table 5.6). F1-scores and accuracy differ by less than 0.03 (3%), with most within 0.001 (0.1%) difference. Since they follow the same patterns by condition, we discuss them in terms of accuracy outcomes for comparability to other multiclass affective classification literature [5, 114, 115, 153, 155].

5.4.1 Subject-Independent Emotion Classification

Subject-independent classification, LOpO-ALL (chance $\approx 25\%$) was run as a single RF trained on all emotions while LOpO-HALF was built on two RFs trained independently for *excited-depressed* and *stressed-relaxed* respectively. For each LOpO

Table 5.6: Overall classification performance across all test conditions and modality combinations by accuracy and weighted F1-scores.

TEST	DESCRIPTION	CHANCE	ACC	F1
LOpO-ALL	Predict one of four emotions	25.0%	34.5%	0.318
LOpO-HALF	Predict one of two emotions	50.0%	58.0%	0.574

level, classification was performed at each window size and gap condition.

Some participants fit the model well, most performed at chance, and, interestingly, a few consistently contradicted the generalized model. For all LOpO levels, window sizes, and gap conditions, **accuracy was very near chance** (Table 5.6, LOpO-ALL and LOpO-HALF).

5.4.2 Participant Classification

Previous results have demonstrated that participants have a *touch signature*: ways or styles of touching which can be sufficiently idiosyncratic to identify the toucher [41, 87]. Individual touch behaviours were both internally consistent and externally unique.

To see if this was true of our data, we performed 20-fold CV ⁷ on the full set of data instances, to predict subject label (*who* performed the gesture) on touch instances, resulting in a classification accuracy of 78%, where chance is 1/30 or 3.33%. High accuracy rates on participant prediction confirms that individual differences are indeed highly expressed in this type of behavioural data.

5.4.3 Subject-Dependent Emotion Classification

With participant classification (Section 5.4.2), we looked for touch behaviour high in both individual differences and consistency. With emotion classification we seek *commonalities* in touch behaviours across individuals, under given emotional conditions. We expect one of the following to be true: (a) participants feeling the same emotions touch the robot similarly, s.t. we can differentiate solely on emotion condition; (b) given knowledge of a participant, we can differentiate between two

⁷Fold-count chosen to balance the reduced data set of subject-dependent models (roughly 1/16 of subject-independent data). Lower folds creates more variance in performance metrics.

emotion tasks; or (c) some combination where a system does not explicitly know who a participant is, but can differentiate given a touch signature characteristic of a specific participant.

(A) is unsupported based on our LOpO results where named emotions are recognized at near chance. We focus this section on the feasibility of personalized models of emotional touch: the consequences of (b) and (c); the effect of noisy or inconsistent data to simulate real-world operation; and finally, how the relative contribution of touch and gaze compare with respect to classification accuracy.

We review classification performance with respect to data factors described in Table 5.4.

Accuracy by Emotion

We break down the average accuracy rates for emotion classification and compare how the classification task affected performance for each emotion (see Fig 5.3).

Unsurprisingly, subject-dependent CV results in higher performance than subject-independent LOpO; notably, however, *Excited* behaviours can be classified at roughly similar rates. There are a few contributing factors to be considered: (1) *Excited* behaviours were of consistently high arousal with quick motions; while *Stressed* was also high arousal, participants often associated it with fighting *Depressed* feelings. (2) Participants provided longer samples of *Depressed* and *Excited* expressions, which led to more data instances when cut into equal-length windows (see Table 5.3).

Window size and Gapping

Comparing classification accuracy by window size, we see that overall, increasing window size improves performance.

We imposed data gaps to simulate real-world loss, reducing temporal interdependency. Where data was uninterrupted (Figure 5.2c,d), classification rates are relatively stable regardless of window size.

While introducing gaps (data discontinuity) causes expected dips in performance, larger window sizes suffer disproportionately. Closer inspection reveals that this accuracy drop-off coincides with a decrease of training instances – most

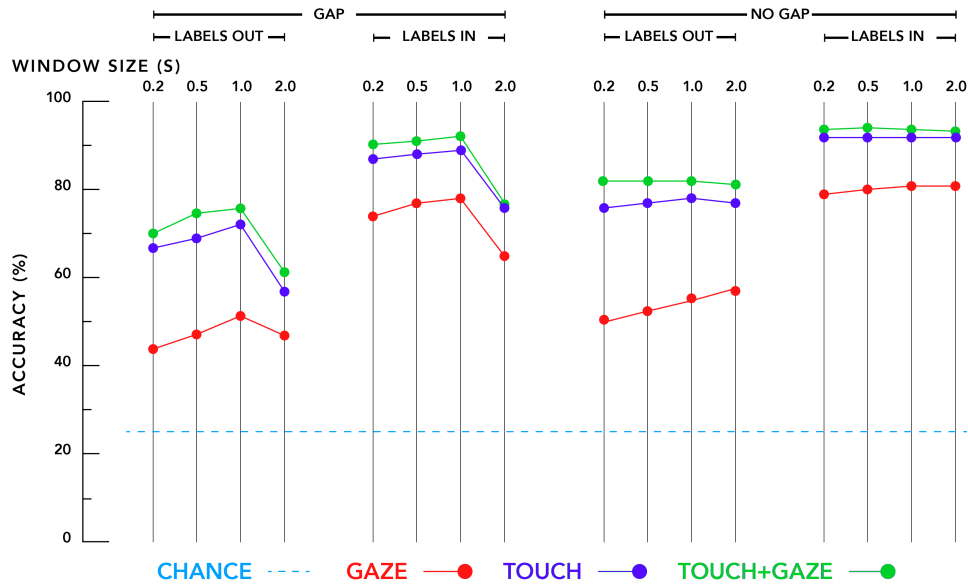


Figure 5.2: Emotion classification accuracy rates from 20 fold cross-validation by *modality* (Touch + Gaze, Touch only, and Gaze only), *window size* (0.2s, 0.5s, 1s, 2s), as weighted averages from Table 5.5. Comparisons are also made between having participant labels included (b) & (d) vs excluded (a) & (c), and where 2s gaps are imposed to simulate data loss (a) & (b) vs no gaps (c) & (d). Including biometric data consistently achieves 90-100% accuracy across windows, labels, and gaps (accuracy dips only under the sparsest data conditions: gapped-2s window cases, regardless of whether subject labels are present).

severely at 2s, where data instance count drops from 7435 instances down to 676, an over 90% data loss.

Participant knowledge

Where participant labels are known (Figure 5.2b,d), classification accuracy improves over cases with no participant knowledge (a,c). This effect is seen consistently across modalities with jumps as high as 10-20% for touch- and gaze-only, respectively.

SUBJECT DEPENDENT PTPT LABELS OUT					SUBJECT DEPENDENT PTPT LABELS IN				
excited	depressed	stressed	relaxed	<--- classified as	excited	depressed	stressed	relaxed	<--- classified as
27.99	2.71	1.21	0.47	excited	30.42	1.96	0.01	0.01	excited
4.70	25.41	1.11	0.37	depressed	2.71	28.87	0.01	0.00	depressed
1.97	1.93	14.63	0.84	stressed	0.04	0.02	18.43	0.86	stressed
1.46	1.14	1.20	12.86	relaxed	0.04	0.01	1.22	15.40	relaxed
(a)					(b)				
SUBJECT INDEPENDENT on 2 RF					SUBJECT INDEPENDENT on 1 RF				
excited	depressed	stressed	relaxed	<--- classified as	excited	depressed	stressed	relaxed	<--- classified as
22.30	10.08			excited	20.41	6.16	3.96	1.86	excited
15.01	16.57			depressed	11.22	10.46	6.90	3.00	depressed
		13.00	6.36	stressed	6.80	8.05	3.02	1.50	stressed
		10.56	6.10	relaxed	4.63	9.22	2.18	0.63	relaxed
LOpO-HALF					LOpO-ALL				
(c)					(d)				

Figure 5.3: Comparing how each classification task performed by emotion using touch and gaze features. For subject independent analysis (c) we trained 2 RFs—trained on *Excited-Depressed* and *Stressed-Relaxed* separately (no between-set classification – blank entry for *Depressed-Stressed*). In contrast, a single RF was trained on all 4 emotions in (d).

Comparing modalities

We refer to Table 5.5 to assess how touch (T), gaze (G), touch + gaze (TG), and biometrics (B) compare in subject-dependent emotion classification performance (20-fold CV).

Taking modalities alone, we see that gaze performs comparatively lower than touch. When participant labels are available (Figure 5.2b,d), classification on both single modalities improve. However, combining touch and gaze further increases accuracy. Particularly under the best condition of maximal information ((d) – with participant labels, no gaps), touch and gaze together can approach that of biometrics performance (97-100%) – in line with previous work showing high classification performance on physiological data [155].

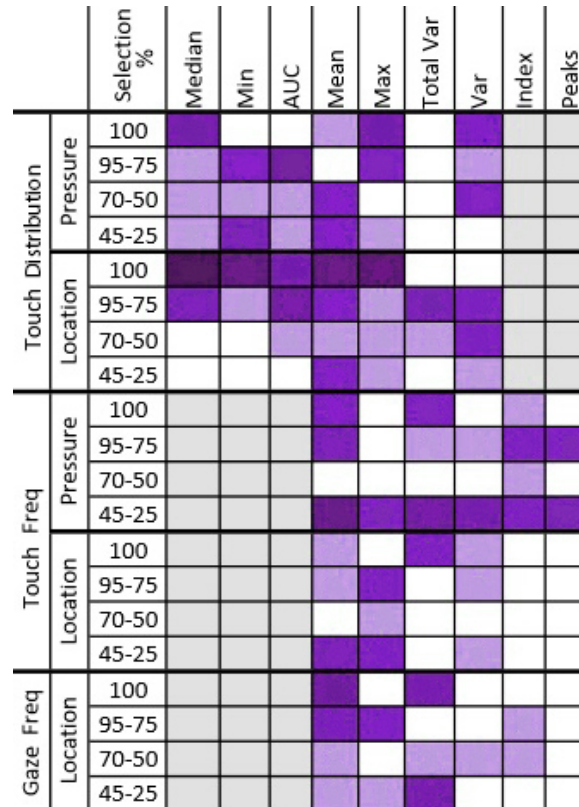


Figure 5.4: Feature selection count by statistic as ranked by Weka’s Best First Attribute Evaluator. Selection % represents how often the feature is selected for use in 100 iterations of 20-fold CV. The dark box for Touch Distribution-Location x Median indicates that this feature is selected 100% of the time; white boxes indicate features that were never selected.

5.4.4 Feature Set Analysis

To understand feature contribution, we ran Weka’s Best First Attribute Evaluator [315] on the Touch and Gaze feature set. This tool iteratively selects the best feature subset for each classification trial in 20-fold CV, producing a list of features and the frequency with which they are selected.

Figure 5.4 breaks down each parameter by modality and relative selection count as a heat map, where each cell represents the number of features of a statistical type

selected at each iteration. Higher saturation indicates a higher number of times selected at this percentage. For example, *Median-Touch Location* was selected in every CV trial.

The most selected features were the 11 calculated medians of touch location, chosen 100% of the time during 20-fold CV. Overall, when using *Classic Touch Location* data, we recommend calculating *Median*, *Min*, *AUC*, *Mean*, *Max* features; in contrast, when using *Classic Touch Pressure* data, *Total Variance* is not chosen at all and may be left out.

5.4.5 Reports of Experienced Emotion

Participants reported their current emotional state with Russell's 2D affect grid [237] during two neutralization tasks and following two emotion tasks. After completion of all emotion tasks, we interviewed our participants on their experience; highlights are covered in this section.

Self-reported emotion movement: In Figure 5.5, there is variation where we expected participants to report emotion movement towards the quadrant extremes. In decreasing order: *Excited* (all 14 participants reported moving towards the quadrant extrema); *Stressed* (13/15); *Depressed* (6/15); and *Relaxed* (2/16). In paired t-tests, we found significant differences in self-reports between neutral and emotion tasks for each of *Stressed*, *Depressed* and *Excited* in both arousal and valence ($p < 0.05$).

Paired t-tests showed no significant difference ($p > 0.05$) in neutralization tasks, nor order effect in emotion tasks.

Figure 5.5 plots each participant's emotion trajectory across the 2D affect grid for each relived emotion instance, from starting state to recall conclusion. Both high arousal emotions (*Excited*, *Stressed*) were consistent with expectations where participants reported a shift in emotion toward the grid corner of the target emotion word.

Authenticity: Each participant self-reported how authentically they experienced the target affect in each emotion task. On a scale of 1–10 with 1 being *completely contrived or artificial* and 10 being *completely authentic as in the original experience*, participants rated authenticity highly (between $7.5 \leq \mu \leq 8.29$) with *Relaxed*

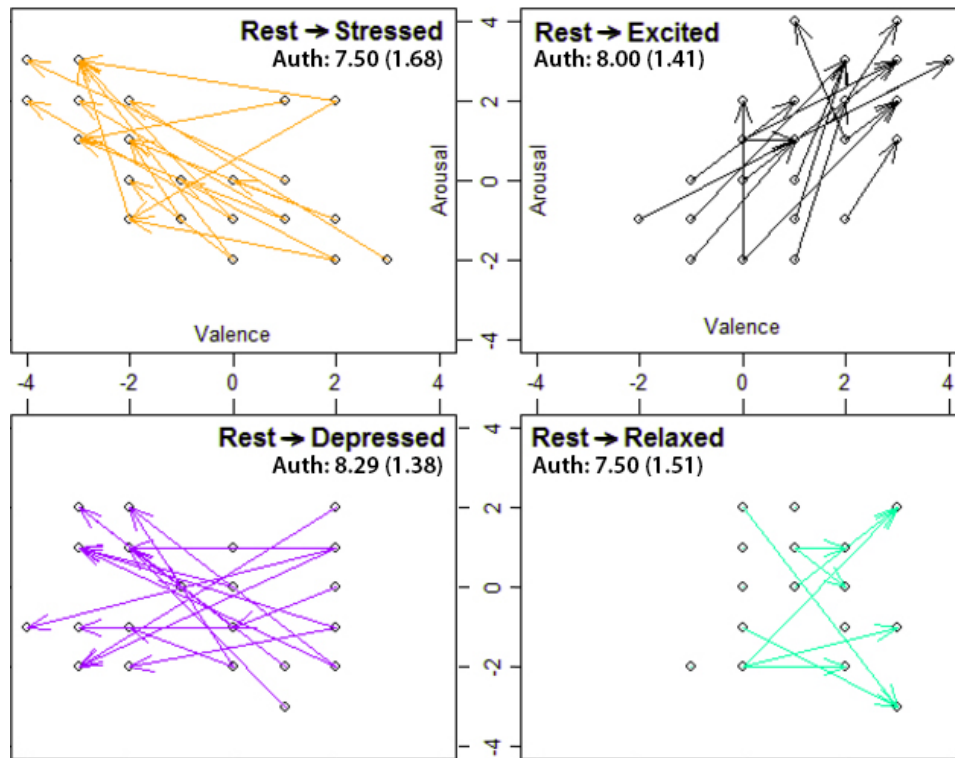


Figure 5.5: Changes in individual’s self-report of emotion after Neutralization (start) and Emotion tasks (finish); N=14 for *Stressed* & *Relaxed* and N=16 for *Depressed* & *Excited*. Overall, we see a move from the origin to the representative quadrant. *Stressed* and *Excited* show the strongest overall change along both **Arousal** and **Valence** axes. *Relaxed* shows the least change with disconnected points referring to “no change” from neutral state.

and *Stressed* tied, and then *Excited* and *Depressed* in increasing order.

Added insight from interviews: For some, immediacy or recency of recalled events helped to highlight emotions. This experiment was run around final-exam and holiday reunion time. Both are cited as reasons for ease of recall.

“I’m leaving to see my family for the first time in three years, I can’t stop being *Excited*.” – P09

“*Excited* was easy – the situation was more recent and was more im-

portant [than my *Depressed* memory].” – P22

“I have a lot of school assignments right now and I kind of toggled between many memories [*Stressed*]. It was hard to pick one to feel but I think that might have added to the feeling.” – P21

“[W]hen I was doing *Stressed*, I felt like I wanted to punch something it was so gut-wrenching.” – P29

The low arousal emotions, *Relaxed* and *Depressed*, moved as expected in valence but not arousal, which remained overall at its neutral “resting” position. In the case of *Relaxed*, this might be explained by perceived similarity between this emotion task and the ‘resting’ start condition.

“*Relaxed* was easy to express because it’s pleasant and I want to feel it and also, I’m sitting on a couch which helps.” – P27; similar reports by P02, P18

For these two emotions, some participants reported that the emotion *Depressed* was linked to *Stressed* in their memories (e.g., feeling stress about exams was also depressing), which may explain some of the unexpected movement in arousal for *Depressed*. Four participants also reported feelings so strong that their *Depressed* memory evoked active tears, while others indicated that these feelings were somewhat mitigated by the experience of stroking a soft body.

“My [*Depressed*] memory was very clear and I was able to recall a lot of details. It really helped to be touching a soft thing and felt like it was taking some of my sadness.” – P26, also P15, P24

Another possibility for both of these emotion targets is that participants were simply unable to turn down their arousal state to this degree during the short time of the session.

5.5 Discussion

We summarize result highlights before contextualizing them in our research questions:

- Using both touch and gaze **improved** accuracy rates over touch alone.

- Increasing window size **had little effect** on accuracy.
- Adding data blackouts or gaps **did not noticeably decrease classification accuracy** except for 2s windows.
- Due to individual differences in touch behaviour, it is necessary to include participants in the training set for potentially usable recognition:
 1. Classification accuracy for *whom* (participant performed a data instance was comparable to that of WHAT (emotion), implying that individual differences can be captured;
 2. Both LOpO-ALL and LOpO-HALF analyses performed at or near chance;
 3. Including participant information in the training set **improved** accuracy rates, but **participant labeling is not necessary** for recognition.

5.5.1 RQ1: Ability of Touch and Gaze to Predict Emotion

As anticipated, accuracy of distinguishing between emotions based on a full suite of biometric signals approached 100% in the best-case model (Figure 5.2a) trained on participant-labeled data (Table 5.5, column *B*). When full-suite biometric signals can be effectively employed, they will give the best result. Even partial biometric sources – e.g., heart rate variability (BVP) alone – do well relative to each of the less intrusive modalities. We can expect improvement in the wearability or embeddability of some biometric channels, so this result is important to note.

Of modalities not requiring sensors to be worn (touch, gaze), touch reaches 92% accuracy⁸, improved with gaze to 94%; however, performance worsens in more adverse conditions. This level of classification accuracy may be adequate for many applications, e.g., when the goal is simply to establish large-scale movement between quadrants.

Classification accuracy favours pressure-location distribution features: At 54Hz, touch distribution features of pressure and location were most frequently selected for emotion-classification performance (Figure 5.4).

Touch frequency-domain features have been used successfully [5]; the contrast may be our relatively low sample rate coupled with short windows (0.2-2s vs

⁸While classifiers differentiated four emotions, each participant performed only two. Chance is thus more like 50% when participant labels are known.

8s in [5]). Further, emotion classification using gaze data appears to consistently benefit from inclusion of features calculated on Fourier transforms of gaze position. Since frequency-domain features are relatively compute-intensive (realtime FFT vs. pre-processable pressure-location set), it may be reasonable to reduce the feature set to touch distribution features where efficiency is a priority.

5.5.2 RQ2: Individuality

Recognition rates increase with greater participant knowledge: LOP results near chance (for both iterations—ALL and HALF) imply low generalizability of a model to other individuals’ emotional behaviour.

Participant knowledge matters, but not labels: We propose a touch-centric robot that exploits individual differences and, instead of an out-of-the-box general training model, builds personalized models of a short list of users. Having participant knowledge is important for classification; all expected users of a single robot should be included in a model’s training pool. However, including participant labels adds only minor benefit (Table 5.5 with labels vs. without) when training data already includes the test participant. This may be due to the relatively high participant classification rate (Table 5.5; chance 3.3%) wherein participant-specific behaviours may influence classification such that even though participants are unlabelled, the system is able to guess. When high accuracy is needed, *a priori* user identification (participant-labelled data) may be a helpful refinement.

Excited is most recognizable emotion: Based on confusion matrices describing per emotion performance (Figure 5.3), *Excited* may be most generally recognizable. The emotion self-report (Figure 5.5) shows that *Excited* was experienced consistently (all participants reported the expected emotion direction). Similar emotional experiences may translate to common touch and gaze expressions in these high-arousal, high-valence emotion spaces.

5.5.3 RQ3: Sample Density for Realtime Responsiveness

Larger windows and including gapped data reduces classification accuracy: With post-hoc classifications, increasing window sizes and eliminating data segments (discontinuities with gapping) reduces data instance count. We discuss the

effects from conditions where greatest data instance count are in no gap-0.2s window conditions and least with 2s gap-2s windows, with respect to real-time classification.

Size: From Figure 5.2, increasing window size from 0.2s to 2s results in marginal improvement of classification under no-gap conditions. In this case, increasing system response rate (by using 0.2s windows rather than 2s of data) may be favourable as little accuracy loss is experienced.

Continuity: Gapping data does indeed drop accuracy by 10% in T , G , and TG (Table 5.5). We considered the possibility that the performance decrease is related to low data instance count, but even when removing that confound and comparing equal instance intervals of gapped vs. non-gapped signals⁹ we found that each single modality’s performance on adjacent data streams (non-gapped) resulted in higher accuracy rates than that of gapped data¹⁰.

Interestingly, for most window sizes (0.2s, 0.5s, and 1s — where gapped and ungapped instance counts are on the same order of magnitude) results suggest data loss should not be devastating to real-time emotion classification of touch, even when the gap (2s) is 10x that of the collected instance (0.2s). Given a relatively predictable signal interruption pattern, we can select a window size range knowing that even if a signal is lost for up to 10x that of the collected window, classification accuracy may still be tolerable.

This performance differential exposes a role of signal continuity in these channels’ expression of human behavior and emotion reaction: a possible explanation is that emotion expression evolves in even short timeframes. While larger, adjacent windows may marginally improve classification accuracies for short (single-window) snapshots, they may introduce error for longer interactions. Periodic system re-training may help to build a more robust user model. Since this may interfere with actual system use, re-training could be suggested as participant behaviour changes and participant classification accuracy drops — an indication of significant behavioural departure from the current model.

⁹Addition of gaps between 2s windows reduces the data set instance count by over 90% (7435 to 676 instances).

¹⁰2s windows / unlabelled participants generated for T : 90.4% (adjacent) vs. 56.7% (gapped); TG : improved to 78.8% vs. 47.5%.

5.5.4 RQ4: Experimental Methodology

We chose an experimental approach based on the use case of a robot pet. Several elements were nonstandard: emotion elicitation method, choice of emotions investigated, study framing (including how existing emotion models may influence the emotion task: a participant interacting with an unresponsive furry object), and analysis aspects. With results in hand, we critique these innovations.

Emotion elicitation: While the technique of memory re-telling was validated by literature [57, 179], we elicited stronger emotional reactions than we expected. In some cases, this could be due to participants playing a ‘good-subject role’, trying to please experimenters [203] and artificially inflating the perceived efficacy of this protocol. However, we anticipate some degree of this characteristic in any laboratory study. Furthermore, we noted some strong physical and embodied emotional reactions (such as genuine tears) that suggests this method could still be a valuable tool, particularly in a laboratory setting where people may otherwise find it hard to act naturally. We plan to employ variations in our own future studies.

Emotion set: We reported both high and low classification accuracy rates, but nevertheless question whether accuracy is an indicator of a successful emotion model, even when corroborated by F1-scores. There is certainly value in accuracy metrics, but underlying assumptions of both dimensional and discrete emotion models present known problems for classification. Specifically, discrete systems based on dimensional models suffer from a problem of distinguishability in which semantically dissimilar emotional labels are placed in the same bins [40].

Study and Emotion Task Framing: We assumed that participants express a roughly *steady state* emotion, felt across the entire memory recall. However, it is possible that strong emotions may be felt only for an instant before autonomic emotion regulation or coping mechanisms take over [101]. The horizon over which we sample a participant’s emotional state, and the assumption of immediacy impact decisions an interactive system should implement. Our discrete classification system can identify differences in minute-long interactions, but cannot estimate an emotional inflection point (i.e., transition from one emotion to another). A truly interactive system would need to react to the *change* in an emotional state and adapt over many samples.

Furthermore, in natural emotional exchanges, interactions with pets or friends allow for error correction: an initial misjudgement can be corrected with further context. An adaptive rather than prescriptive model might go further towards develop a meaningful relationship over a direct and immediate call-and-response instructing interaction [264]. Using touch data in context with gaze and biometric analysis lays the groundwork for extending haptic human-robot interactions from instructional directives to meaningful conversational relationships.

5.5.5 Implications for Social Robot Applications

From our findings, we consider next steps in designing the classification system for our social touch-centric robot.

Out of the three nonverbal modalities we studied, touch may be most relevant for applications such as social robot therapy. Our findings indicate that for a previously known user, *distinguishing between a few emotional states is feasible for touch-alone*. This provides intriguing opportunities for development of therapeutic robots that could run human-affect recognition and respond by adjusting their behavior.

While gaze and biometrics improved classification, their use in practical scenarios remains challenging. For robust detection of gaze, the user must always face the robot at a certain angle or wear a calibrated head-mounted gaze tracker. Including biometrics is even more restrictive as participants must don a series of body-hugging sensors, then remain emotionless during periods of neutral user calibration before departures from neutrality (emotion) can be detected in signals such as heart rate and skin conductance. Embedding biometric sensors into the robot system may be possible but still poses some difficulty: touch interaction with the robot typically consists of momentary touch contact that may be too short and infrequent for measuring biometric signals. However, these sensory systems can be integrated in situations with careful sensor placement for gaze attention and training data collection sessions.

To be used effectively in therapy, an expert such as a therapist would need to introduce the robot and guide potential users in providing training data for recognition of emotions via touch. As participant-knowledge appears to be a key component to increasing emotion classification performance, we can conceive of

a system training procedure that extends beyond simply including participant info. The robot could be personalized to first recognize and then work from a custom user profile where accuracy is crucial. Although this implies a setup cost for use, potential benefits in environments where real animals cannot be used (such as some hospital environments) may compensate.

5.6 Conclusions

We presented affect classification results from emotionally influenced touch and gaze behaviours, verified against better-understood biometric data. Participants recalled intense emotional memories spanning Russell’s 2D arousal-valence affect space, namely *Depressed*, *Excited*, *Stressed*, and *Relaxed*. We collected data across the three modalities via a custom fabric touch sensor embedded in a small furry stationary robot; a gaze tracker; and a biometric suite including skin conductance, respiratory rate and heartrate variability. Our data is both quantitative (sensor capture during interaction, and self-ratings of emotion genuineness and intensity) and qualitative (post-experience interviews).

For models trained with test participant data using pressure-location features, the overall emotion recognition rate was roughly 83% for touch, 87% for touch + gaze, and 99% for touch + gaze + biometrics. Performance drops steeply when test participants were left out of the training model, resulting in 31%, 31%, and 29%, approaching chance (25%). We tried increasing the feature set by incorporating frequency features for touch and gaze modalities. This resulted in emotion recognition rates of 79% for touch frequency features, 85% for frequency and pressure-location touch features, and 85% for touch frequency, touch pressure-location, and gaze frequency features combined. LOpO performed similarly poorly at 30%, 32%, and 35% respectively.

We summarize findings that will inform our next stage of design for robots capable of real-time emotion classification:

1. Emotional behaviour encoded in touch and gaze interaction may be sufficient. While including biometric data greatly improves accuracy, current technology requires they be worn, resulting in a more restrained experience. Setup interferes with natural emotional expression and sensors affixed to the hand and

body can feel restrictive.

2. An individualized training or calibration phase is crucial for a personalized prediction system. Increasing participant information greatly improves the classification model's prediction accuracy. While this stage likely requires guidance from an expert or therapist, the training investment facilitates the learning of user-specific characteristics and develops a more robust user behaviour model, thereby allowing for a personalized and productive experience.

3. Sampling density and feature count may be reduced to improve computation load. During real-use, the speed of classification and reaction is a serious concern. Lossless continuous capture is ideal, however, in real-time we may find that packets must be dropped from slow or problematic data captures. We experimented with introducing gaps in data for this reason, and our findings indicate that interruptions in data collection at up to 2s intervals may be tolerable.

4. Limitations of commonly used emotion models should inform future research in this field. Although we achieved possibly usable classification rates, reflections from the field suggest that existing affect models have clear limitations that must be addressed [38]. People do not experience emotions in isolation nor discretely; emotional experiences follow a trajectory with distinctive peaks and valleys. Future detection systems must model the rise and resolution of an experience. While this study used a stationary robot, a deployed interactive system must acknowledge that its response has influence over user emotional reaction, necessitating dynamic adjustments to behaviour modelling.

Part II

Dynamic Emotion Modelling

In Part II, we iteratively refine our computational model of emotion expression. We begin by proposing a multistage emotion elicitation and self-reporting protocol (Chapter 6) and conclude with a reflection on how to advance design on technology that supports human emotional experiences (Chapter 9). Building from our experience with the work in Chapter 5, participant comments, the literature on emotion regulation, and from our own personal lived experience, we acknowledge that memories focusing on a single strong emotion inevitably involve resolution, personal background, and event context that complicate the emotion experience. To highlight the richness and dynamism that is inextricable from the emotion recall experience, we developed a multistage emotion labelling protocol that allows us to root the emotion data in personal history. This protocol is presented as published in Chapter 6.

Developed in tandem with the contributions in Chapter 6, Chapter 7 is the implementation follow-up. While the former focused on protocol development, describing the data labelling process and justifying / validating the procedure, the latter features the analysis of the multimodal data and describes important conclusions coming from incidental touch (as collected by force-sensitive resistors).

To highlight the practical implications of creating personalized models for emotion classification of affect expression, we constructed a proof-of-concept of the model training process. In Chapter 8, we outline how we collected both biosignals (heart rate and skin conductivity) and touch behaviour expressed during an emotional storytelling task. By labelling this emotion data with time-varying emotion in multiple stages, we constructed training and test sets for emotion classification.

Finally, in Chapter 9, we look forward to the development of emotionally responsive devices and consider the guiding principles that can best serve the betterment of their human users.

Chapter 6

Dynamic Emotion Detection: A Multistage Emotion Self-Report Labelling Protocol

Summary

Many emotion classification and prediction approaches focus on emotion *state*, defined as static and single-valued. In contrast, our in-body experience is of sensations that can quickly evolve, consistent with scientific evidence of physiological regulation mechanisms. Can we reframe classification to estimate dynamic emotion parameters at interactive rates?

For insight into dynamic emotion characteristics, we developed a multipass labelling protocol to capture controlled yet genuine emotion evolution elicited as 16 participants played a tense video game. We analyze and align multiple self-report outputs, inspect the signals for emotion dynamics, and consider label metaphors of **position** and **angle** – “where I am” vs. “where I’m going”. Finally, we reflect on the benefits and drawbacks of such a protocol for developing models of fast-evolving emotion.

6.1 Introduction

Whether building robots that detect anxiety through touch interaction or video games that dynamically adjust level difficulty to optimize player engagement, computational models of authentically developing emotions are the foundation of technology. Challenges arise in developing these computational models from true and spontaneously evolving emotions.

Emotion theorists have long observed time-varying dynamics of emotion expression, attributing them to complex neurological and physiological regulation mechanisms [100], appraisal effects [198], cognition and contextual factors [194, 211]. To simplify in-lab research, computational emotion modelling often relies on an “emotions-as-point” metaphor [38, 169], represented as a dimensionless point in an emotion plane in which self-reporting static emotion labels for classification involves easy-to-read scales, often along dimensions of arousal, valence, and dominance [42]. While these models are convenient, for realtime use we need to recognize emotion evolution over time, rather than distilling a lengthy event into a single label.

Going from theoretical to computable: Obtaining authentic emotion data is a significant obstacle. Our memories and emotional assessments are affected by time and reflection [198, 211]; how representative can a reporting scheme be of someone’s “reality”? Commonly used labels on the arousal-valence circumplex model [237] or PANAS [309] or SAM [30] (among others) quickly become intractable for sampling at the rates at which emotion can potentially evolve.

Emotion is personal: Independent of the measurement instrument, self-report of emotion incites questions of generalizability across the population. A researcher’s understanding of the instrument scale may be very different from that of a participant [38]; our comprehension of an emotional ‘landscape’ or internalized emotion frames of reference are highly subjective, influenced by life experiences and personal history [20]. We presume that any set of ground-truth labels for self-reported emotion are similarly personalized: i.e., the experience or scale for *anger* for one person may not be recognizable for another.

We propose that evaluating emotion based on dynamic qualities will advance the accuracy of machine recognition of human emotion experiences. Better fore-

casting of a user’s near-future emotional expression allows for system responses that are temporally and situationally appropriate.

6.1.1 Approach

We assess the viability of building computational emotion models based on **dynamic** conceptualizations of emotion change to bolster our capacity to predict and respond to human emotion based on observed behavior or self-reports. Multi-pass labelling requires high investment in early model building, which can pay off by highlighting how to optimize labeling in later real-time use. As outlined in Figure 6.1, this paper evaluates reporting consistency between passes of a data collection and emotion labelling methodology, leaving model building and classification performance for future work.

Specifically, we reflect on our multi-pass protocol which (a) **triangulates** emotion self-reports with modality-agnostic observable data; and (b) employs co-creation of **personalized** calibrated emotion scales which form the frame of reference for multi-pass self-reports, collected with minimal intrusion on the primary emotion event. Using a joystick for spatiotemporally high-resolution post-hoc ratings, we can construct data windows that are (c) **versatile** to accommodate a variety of emotion metaphors at our choice of time scale.

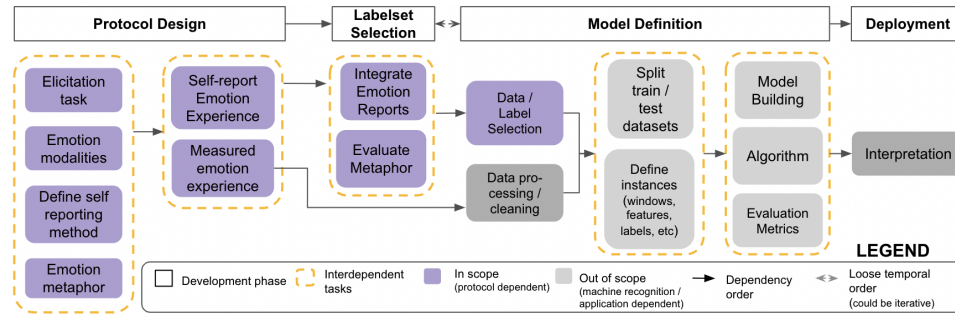


Figure 6.1: Roadmap for developing an emotion-prediction engine for an emotionally responsive application.

6.1.2 Research Questions

Two lines of inquiry guided our assessment of this approach to operationalize real-time models for emotion forecasting.

1. Do user-centered emotion reports add new information? Nuances in users' emotion language, manifesting as apparent inconsistency, can interfere with emotion model performance and validity [148]. We center users by including (a) personally calibrated emotion scales where we create a shared understanding of instruments and measures [38], and (b) multiple labelling passes at different resolutions and retrospective distance; then assess the information gained from these elements. For example, do people rank common emotion words similarly? In what ways does labelling data differ by pass? What do we gain from quantifying the differences?

2. How might we incorporate the dynamic nature of emotions into our computational models? Operationalizing dynamic emotions requires models that represent the natural evolution of an emotional experience.

We begin with the prevalent movement-based metaphor of *emotions-as-position* ('where I am', an ordinal value on an emotion scale), and propose another of *emotions-as-angle* ('where I'm going', the direction and sharpness of change). We add to these previously-proposed emotion dynamic measures of inertia, instability, and variability [130], and compare the properties of each with each other and in between-participant variability for insights into how they might have value for responsive computational models.

Through these investigations, we contribute:

- A multipass labelling protocol with insights into how to employ triangulated emotion labels, including the role of personalized emotion word calibration;
- Insights into the descriptive properties of various dynamic emotion parameters, relating to their potential for use in responsive computed models.

In the following, we root our protocol development in the existing literature, describe the devices and instruments we created to measure continuous dynamic emotion, outline the data collection procedure, and evaluate the data according to our questions. In discussing our findings, we consider where these new model elements may provide the greatest value.

6.2 Related Work

Protocols featuring internally consistent emotion metaphors, measurement instruments, and elicitation procedures increase the likelihood of representing true participant experiences [38].

6.2.1 Emotion Self-Report

Classifying emotion requires capturing and labelling emotional experiences. Representation thus impacts how we ask users to report their experience.

Russell’s circumplex model [237] is a commonly used instrument depicted as a spatially continuous 2D space of arousal and valence (plus dominance in 3D [13]). It underlies popular labelling schemes, most involving a participant locating emotion words on its axes; e.g., words associated with PANAS, the Positive-Negative Affect Schedule [309].

The Self-Assessment Manikin (SAM [30]) makes this more natural with Likert scale dimensions [274, 317].

Natural language reporting methods are used when experiences (maybe a self-contained memory [41], or a touch [114]) are sufficiently brief, simple to fit a single label, and precede an opportunity for the participant to report without experiential interference. They become intractable for segments that are longer than a few moments, span multiple emotions, and/or require rapid computed response (before the segment ends).

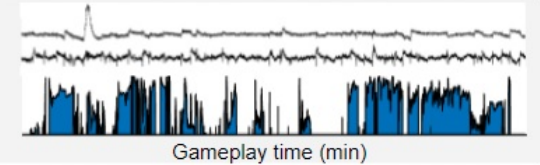
Still with a dimensional representation, others have collected *temporally continuous* emotion ratings using a mouse- [61] or a joystick [263, 318]. For hands-free activities, a joystick allows for high temporal-resolution concurrent reporting, but at the cost of emotional intrusiveness. Post-hoc ratings require review of a recorded experience.

We drew on these approaches to design our own **joystick-based continuous emotion annotation** system.

Participant Task Details in Order of Performance

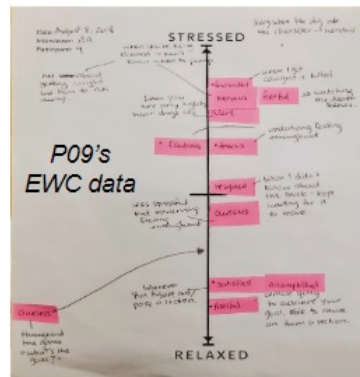
Data Representation

1. Primary Emotion Activity (PEA): Participant plays a video game to elicit authentic emotion, to be recorded as classifiable physiological data.
Data: emotion encoded data; emotion task timeline; video recording of participant performing the video game.



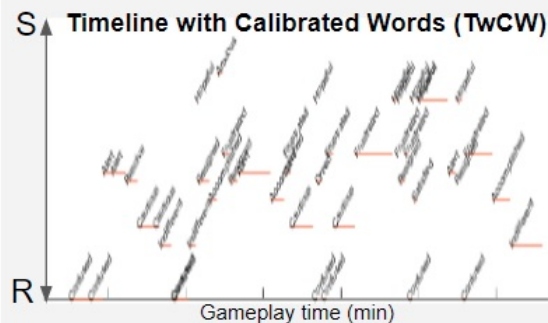
Calibration

2. Emotion Word Calibration (EWC) - Participant places common emotion words from a list on a scale from [-10 Relaxed] to [10 Stressed]
Data: Participant-specific set of Calibrated Words, labelled with distance from "Relaxed"



Game Review 1

3. Calibrated Interview (CI) → TwCW: researcher and participants watch PEA video and together mark notable moments on (Relaxed ↔ Stressed) scale, including emotional extrema and other standouts. Participant compares their word scaling here with those in Task 1.
Data: Interview transcript, annotated emotion word timeline as a **TwCW**.



Game Review 2

4. Continuous Annotation (CA) - participant watches video of gameplay and uses a custom joystick that holds position (rather than return to center) to provide continuous annotation of their recollected emotional experience on Relaxed ↔ Stressed scale
Data: continuous emotion time-series on Relaxed-Stressed scale.

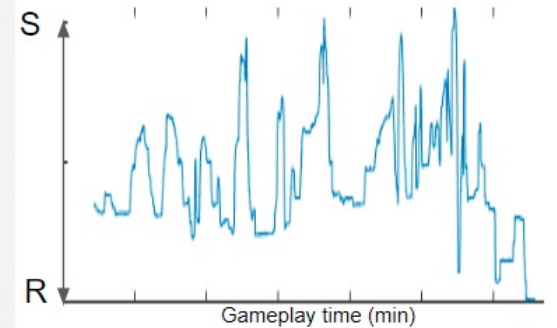


Figure 6.2: Participant tasks and resulting data. At lower left is an EWC example: word stickers placed on a Relaxed-Stressed scale, plus P09's other annotations. The latter resulted from P09 later contextualizing their in-game experience.

6.2.2 Characteristics of Emotion Dynamics

The methods above imply emotion as “state”. Even models that feature sequences (i.e. Bayesian emotion models) denote each stage as a single state [59, 211].

Regarding emotion instead as a *process*, as in appraisal theory [198], may better reflect human experience; but this perspective must be operationalized. One approach is to calculate *emotion dynamics*, by quantifying progression in three fluctuation parameters on one’s emotional movement: (1) *inertia* (the time it takes), (2) *instability* (by how much), and (3) *variability* (the range of those changes), calculated as autocorrelation, mean square of successive differences, and within-subject variance respectively [130, 278]. Using these summary metrics over a report time series, researchers have evaluated emotional character arcs in movies [121], examined the role of exercise in emotion regulation [23], and even predicted mood disorders [278].

Can we use these markers at high resolution, to capture transitions and support concurrent response or are other motion characteristics more appropriate? We investigate **sourcing labels from a report’s emotion dynamics**.

6.2.3 Labelling and Timing

Timing is key to regulation, reflection, reporting, and in-event reactivity. Emotions evolve at multiple time scales; an event may evoke a different emotion after cognitive reflection on an in-time reaction [211]. The optimal timing for capturing a self-report is complex. Too soon may curtail rich and valuable reflection [198]; too late incurs memory decay [76]. Concurrent emotion evaluation is typically impractical: probing for labels is intrusive and distracting – naming a feeling is a form of reflection and regulation [100, 102, 103].

To capture reflection and generate training data for future responsive models, **we collect reports in two passes and use multi-timescale labelling** – giving time for self-reflection, and mitigating memory degradation with video reminders.

6.2.4 Emotion Elicitation

Where applications require in-time recognition of emotion, data must represent realistic emotion expression [93, 127]. Relived or recalled emotion is one proxy [42,

80]. Participants are prompted with an emotion word (the single label) and asked to recount the story of a past intense experience.

While successful in eliciting authentic and wide-ranging responses, this oversimplifies an episode to emotive homogeneity [42]. Furthermore, participant stories are hyper-individualistic, not amenable to a search for commonalities. Conversely, entertainment media can root participants in a more uniform elicitation stimulus, with many validated video and music clips used successfully for this purpose [57, 103]. Video games have shown promise in producing physiological responses analogous to that of real life evocations [292].

Here, **we use a horror video game to elicit emotion**. This genre has shown high user immersion and engagement, evoking emotions from anxiety to happiness and contentment [214].

6.3 Data Collection Protocol

Our priority was to obtain triangulating data views on the emotional space of momentary transitional experiences. $N=16$ individuals (8 reporting as male, 8 female; 19 to 34 years of age, half under 25) participated. Each participant supplied self-report data that demonstrates our protocol, by completing four tasks as outlined in Figure 6.2 and detailed below.

6.3.1 Participant Task 1: Primary Emotion Activity (PEA)

To demonstrate this protocol, we use video game play to elicit authentic and spontaneous emotion. We chose *Inside* [222] to stimulate anxiety without graphic violence and spark moments of accomplishment or satisfaction, all with easy-to-learn keyboard controls. We selected participants for their affinity for video games, excluding those with experience of *Inside*.

For reviewing the primary gameplay experience in later passes, we videorecorded participants' faces and game screen (OBS¹, 30fps). Gameplay averaged 13:24m (min 8:25, max 21:37, SD 3:88).

¹Open source video recording and streaming. <https://obsproject.com/>

6.3.2 Participant Task 2: Emotion Word Calibration (EWC)

To contextualize individual interpretations in later steps, participants rated up to 15 emotion words, two write-ins and 13 from the PANAS [309]: *Cautious, Satisfied, Hopeful, Frustrated, Anxious, Nervous, Threatened, Resigned, Alert, Accomplished, Fearful, Dread, Curious*. Figure 6.2, lower left shows P09’s sample scale, ordering these words between Relaxed to Stressed (chosen to represent diametrically opposing quadrants from Russell’s circumplex of Arousal vs. Valence [237]).

We measured the distance from the Relaxed line to each word’s placement, mapped it to a 20-point scale ($[-10,10]$), and aligned the words and their scaled heights with the interview (I) transcript via timestamps, to form a time-series of emotion word (and synonym) height.

6.3.3 Participant and Researcher Task 3: Calibrated Interview → Timeline with Calibrated Words

In the first labelling pass, participant and researcher jointly reviewed the gameplay video. The participant indicated emotionally notable points while the researcher marked them on a gameplay timeline. Because participants had previously undergone a word calibration, they were primed to consider how the offered vocabulary were distributed across the emotion scale.

From the Task 2 Interview transcript, we found synonyms and root words using Python’s Natural Language Toolkit [225]. We constructed the **Timeline with Calibrated Words (TwCW)** by placing values where a root matched the EWC, with each value a numerical distance from Relaxed. For example, P09’s comment “*The barking in the distance filled me with anxiety*” would map the calibrated point value of 14-Anxious (synonym of *Anxiety*) on P09’s calibrated 20-pt scale at the timestamp in the game where the dogs began barking.

Participant language included $\mu=37(\sigma=7)$ calibrated word instances with annotation frequency $\mu=0.05(\sigma=0.015)$ words/min; duration was roughly double gameplay.

6.3.4 Participant Task 4: Continuous Annotation (CA)

In the final pass, participants reviewed the PEA video without pause. They used a custom joystick (holds position rather than returning to center) to continuously trace a 1-dimensional emotion rating between predefined extremes (inspired by [61, 263, 318]), here employing the previously calibrated axis. The result is a continuous rating time-series (256Hz) corresponding to the original gameplay, down-sampled for analysis to 30Hz to match the video framerate.

During annotation, smoothed joystick position is graphically rendered as the height of a bar on the video screen, for feedback on proximity to a more Relaxed (blue) or Stressed (pink) emotional moment.

6.3.5 Task Order

Task order was carefully chosen to minimize influence on emotion elicitation while increasing the likelihood that participants would use a common set of emotion words to describe their experience. During Step 3, the interview allowed players to explicitly process their emotions out loud, guided by researchers looking for notable emotional events – strong emotions, startling or uncomfortable moments, odd behaviour etc. Leaving the joystick evaluation as the final step lets participants internalize and contextualize the emotion scale in preparation for the continuous annotation.

6.4 Exploring Multi-Pass Emotion Self-Reports

Our present analytical goal is to explore the properties of and relationships among the reports obtained with this protocol, primarily by examining the degree and nature of their [dis]similarity over a range of metrics, and probing for physical intuition among them.

6.4.1 Commonality in Interpreting Emotion Words

To assess across-participant similarity of calibration ratings (as a proxy for model generalizability), Figure 6.3 plots *rating variance* for each of the calibrated words in order of decreasing agreement (increasing variance).

For a quantitative view of cross-participant consistency, we also conducted an *intra-class correlation (ICC)* (inter-rater reliability test [160]). For the subset of emotion labels rated by all participants (*Anxious*, *Cautious*, *Frustrated* and *Satisfied*), we found $ICC(2, k=16)=0.99$, $p \ll 0.01$ ($\alpha=0.05$, $CI=[0.97, 1.0]$), based on mean rating over an absolute-agreement, 2-way random-effects model. ICC values > 0.9 indicate high reliability [160], suggesting these ratings are overall highly similar across-participant for this set of emotions. Indeed, the four rated by all participants had an $ICC(2, k=16)$ of 0.99.

However, this agreement varies as set size increases, first decreasing monotonically then dropping sharply at *Satisfied - Resigned* to $ICC(2, k=4)=0.83$. This may be partially due to the relative sparseness of ratings.

Taken together, these results support that there are **substantive differences in how individuals interpret emotion words, highlighting the importance of personalized models.**

6.4.2 Self-Report Modality Consistency via Time Series

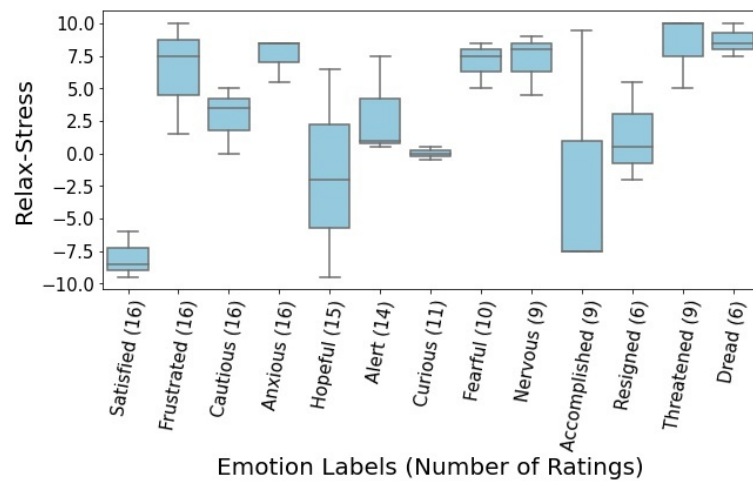
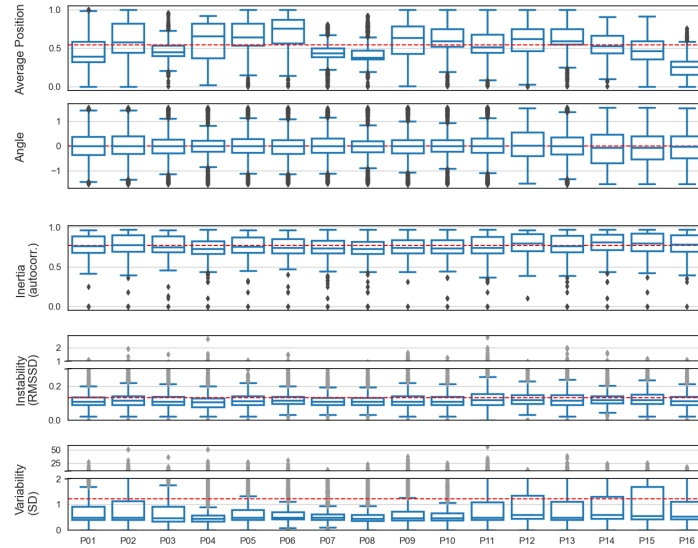
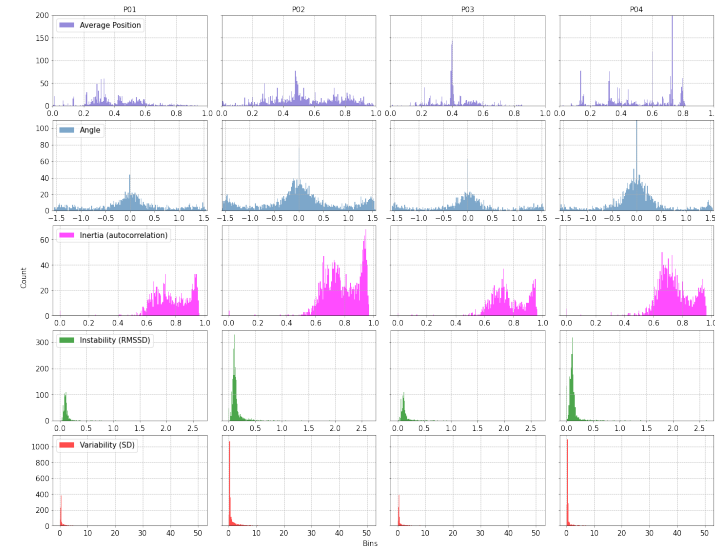


Figure 6.3: Rating variance by calibration word, ordered by number of participants who provided a rating for that word.



(a) Boxplots of emotion dynamics of Continuous Annotation (Task 4) data, by Participant ($N=16$). Position ($M=0.5465$, $SD=0.2221$), Angle (0.0049 , 0.7127), Inertia (0.7666 , 0.1215), Instability (0.1316 , 0.1086) and Variability (1.2165 , 2.4079).



(b) Representative subset of label distributions: emotions-as-position (average position; *purple*), emotions-as-angle (angle; *blue*), Inertia (*magenta*), Instability (*green*), Variability (*red*). Note that longer gameplay results in more samples.

Figure 6.4: Comparison of summary statistics and histograms by emotion parameter.

High similarity between self reports indicates consistency and perhaps interchangeability of report modalities; differences might suggest invalidity of one or both, or that they capture different information. Interpreting within-participant TwCW and CA as time-series, we use standard time-series analysis methods [187] (with appropriate condition verification steps) to check for signal similarity – Pearson’s correlation – and confirm that both data streams are appropriate responses to a common stimulus – Granger’s Causality [258]).

Test Preparation: Using raw report data, we first confirmed that **both time-series were stationary** with the Augmented Dickey-Fuller (ADF) test (Bonferroni-Holm correction $\alpha = 0.05$, $p_{BH} < 0.02^2$), and that their statistical properties did not change over time [92]. Prior to evaluating cross-correlation between the two reports, we verified that each was not auto-correlated to avoid artificially inflated correlations [66]). With Python’s *statsmodels* [251], all peaks were at lag=0 for all participants’ TwCW and CA auto-correlation plots (i.e., both signals present low correlations at all lagged versions of itself). We conclude that **neither signal is self-similar**.

The TwCW and CA self-reports are sampled at different times and resolutions (0.05Hz and 30Hz respectively). We downsampled the CA series rather than interpolate the sparse TwCW, to minimize bias.

Pearson’s Correlation for signal similarity:³ P01, P02, P08, and P14 had moderate correlation coefficients for the two emotion self-reports (CA and TwCW) at $\rho > 0.3$ ($p_{BH} < 0.05$). However, in general there was no significant correlation between the report streams: p-values exceed the threshold after a Bonferroni-Holm’s adjustment to $\alpha = 0.003$. We infer that **individuals’ self-reports differed** in the metrics we observed.

Granger Causality Test for source plausibility: Although Granger cannot confirm direct causality between different variables [270] (i.e., it does not claim TwCW causes the CA values), we employ the test to evaluate whether time-series for CA could *forecast* TwCW and vice versa. We employed a Bonferroni-Holm correction

²For all except P01 (TwCW): $p_{BH} = 0.07$, ADF test statistic = -2.671

³Pearson’s correlation results at $\alpha = 0.05$: P01 ($\rho = 0.38$, $p_{BH} = 0.142$), P02 ($\rho = 0.38$, $p_{BH} = 0.235$), P08 ($\rho = 0.43$, $p_{BH} = 0.235$), P14 ($\rho = 0.37$, $p_{BH} = 0.245$)

($\alpha_{BH}=0.05/N$, N = number of participants). We found significance for 15 of 16 participants ($p_{BH} < 0.048$), suggesting that one label stream could be used to forecast the other for all except P02. This implies **the data streams are appropriate as responses to the same stimulus**.

6.4.3 Comparing Motion Characteristics of Emotion Dynamics

We next examined how various parameters computed on these time series might reveal differing insights. In this scope we included: signal *Position* (the prevalent standard, and following an “emotion-as-state” metaphor); *Angle* (drawing on an alternative metaphor for emotion as directional and changing); and [130]’s three emotion dynamic parameters of *Inertia*, *Instability* and *Variability*. Our investigation included comparing these time series (original and computed) through summary statistics and histograms, all by participant.

Data Preparation: We further analyzed each participant’s Continuous Annotation⁴ data by first partitioning the continuous self-report data into 500ms windows (window count $\mu=1587.75$, $\sigma=462.50$ by participant). Where window boundaries do not coincide with a logged data point, we imputed with the previous data point, turning our time-series into a higher-resolution but stepped signal.

We computed *Position* labels from windows by mean value; and *Angle* labels as the rate-of-change per minute from a least squares linear fit, in the form of an angle $\theta \in [-\pi/2, \pi/2]$. Using R’s *psych* package [229], we calculated *Inertia* (autocorrelation coefficient), *Instability* (Root Mean Square of Successive Differences (RMSSD)) and *Variability* (Standard Deviation (SD)) by window for each participant [130].

Comparing Summary Statistics and Histograms by Parameter: Figure 6.4a shows signal statistics for each participant and parameter. The means for all five measures track closely across participants. However, spread differs: *Inertia* is relatively tight and symmetric, *Variability* is broad and highly asymmetric, *Instability* in between.

⁴Tests for equivalence between the two sets of self-report (CA and TwCW) across each of the three emotion dynamics parameters (two 1-tailed paired samples t-tests [171] per dynamic measure) were inconclusive ($p > 0.5$, $t(15) \ll 0.001$, $d \ll 0.001$). Subsequent emotion dynamics explorations were done on the higher resolution CA data.

In an alternative view, Figure 6.4b shows the same parameters and signals, but now as histogram distributions. Data for these four participants are reasonably representative.

Comparing these two representations of the same underlying data is insightful. For example, while in Figure 6.4a *Position* is clearly less stationary than *Angle*, 6.4b indicates the form that this takes (broader spread, spikiness). And while the dominating feature of the other three ED's boxplots is the uniformity of means across participants, histograms reveal their internal parameters as starkly different: *Inertia* is broad and high-valued, the others low-valued with very long tails.

No insight was gained from visual analysis of spectral qualities (from a Fast Fourier Transform) of all five parameters.

Which is Best? The preceding section's results demonstrate that the relatively high resolution of the CA report (30Hz raw, parameters computed at 2Hz) affords computation of a variety of descriptive parameters. Getting to the root of what the differences in label representation mean will require approaches assisted by synchronized physiological data views.

6.5 Discussion

Compared to past studies of dynamic changes in behaviour or mood [130], our video game task is short and densely reported. With its data we reflect on our questions and protocol, highlighting implications for high-resolution real-time models.

6.5.1 Multi-Pass and Personalized Emotion Reporting

To estimate emotion evolution by-the-second, we can select a single dimensional emotion scale and collect self-reports (as in our CA data). How does adding scale calibration and a review/interview phase enrich this report stream?

Personalized scales clarify what may be generalizable, as well as improving personal models' accuracy: Asking participants to project a set of emotions onto a specified emotion axis grounds the ratings in an individualized experience between the Stressed-Relaxed extremes. Plotting the ratings across commonly used words (as in Figure 6.3), we see that words with low rating variation – *Satisfied* and *Anxious* – may be useful as emotion reference frames. In contrast, high variance

words like *Hopeful* or *Accomplished* may be less useful for labelling without additional interpretation.

Multipass reporting increases label versatility: A continuous annotation of emotion communicates a highly personal experience at a resolution that is otherwise difficult to solicit. As a continuous quantitative signal, we can model emotion as a regression for high-resolution forecasting or elect to discretize (or bin values) for categorical classification. Additionally, we can compose an entirely new time-series by incorporating our personalized scale into an interview as a lower resolution signal where continuous annotation is impractical or unnecessary.

Disagreement may indicate synergy, not conflict: Data from our two passes (annotation and interview) are not correlated enough to be interchangeable, yet causality results indicate they are highly related. Perhaps each has its own authenticity and value, which could be optimized in protocol refinement, then extracted and integrated. Further work is needed to identify the different perspective that each brings.

6.5.2 Incorporating Dynamics into Emotion Models

Reading signal characteristics (like autocorrelation, mean successive differences, variance) as measures of emotion inertia, instability and variability connects them to lived experience. What can they mean for intuitive predictive models?

Momentary emotion dynamics as characteristic, not label: Inertia, instability, and variability can help elucidate “slow emotion” in mood disorders [278], but lose meaning in rapid-response timescales, and thus as emotion labels. Reframed as informative signal statistics, they yield hints such as emotion variability’s larger spread suggesting extra *sensitivity* (Figure. 6.4a) which could inform model development, e.g., by identifying archetypal behaviours for improved model selection.

An abundance of metaphors to fit the need: The metaphor of “emotion-as-position” does not capture “fast” emotion dynamics. For example, *Angle*, which captures relative differences in emotional intensity, has a natural physical meaning of directionality – *where I’m going*, not *where I am*. We have seen that *Inertia* and *Instability* respectively lend insight into responsiveness of emotion to stimuli, and emotive range.

Context may dictate choice of label metaphor. To identify if someone is *Excited*, we may choose a **position** representation; to catch *getting Sadder*, **angle** may work best. A **position** metaphor is more versatile; **angle** can be estimated from a set of points but the reverse requires additional information.

6.5.3 Protocol Reflections

At high temporal resolution, reporting can be intrusive and tedious. We reflect on our multipass labelling procedure for tradeoffs and consider possible improvements.

High-resolution labelling does not have to be intrusive. Since emotion reporting happens before and after elicitation, this labelling protocol accommodates any combination of sensing modalities. The emotion experience can unfold naturally, since labelling is done in review.

High time-resolution may be best for short time-scales. Continuous annotation is great for tracking emotion evolution during a 20-min video game session but onerous for prolonged review; and this protocol’s overhead is unsuitable for occasional low-effort check-ins. Multiple passes are ideal for tasks that promote dynamic emotional experiences over a short time, and where reflection and review-dependent labelling are valuable: e.g., therapeutic activities, recalling a memory, playing a game, interacting with an agent. Simultaneous emotion rating may be possible while watching a video or listening to music: joystick annotation during the elicitation, so long as the elicitation activity is hands-free.

Ordered tasks cannot be counterbalanced. We carefully selected the order of tasks to prioritize emotion reflection and recall. The tradeoff for lightening the mental effort and reducing time investment for multipass labelling means that we cannot counterbalance order for the Calibration (Step 2), Timeline with Calibrated Words (Step 3), and Continuous Annotation (Step 4). We are unable to evaluate generalizability of the labelling passes in other protocol orderings.

6.5.4 Future Directions

This paper is an initial exploration into the labelling procedure for dynamic emotion modelling. We highlight where future directions are highly promising.

Parameters computed on high-resolution data are different. What does this mean? To get behind different characteristics in computable descriptive parameters, one approach is to compare with other high-resolution data streams such as EEG and facial encoding. We plan to do this by focusing analysis on particular events (e.g., timeline regions stimuli known to trigger reactions in all – a scary spot in the game), and see how these parameters look across multiple participants when calibrated in a variety of ways.

At what time scale does calibration change? We calibrated our scales prior to the emotion elicitation task. Could engaging in a highly emotionally charged activity influence the rating scale upon reflection? In future iterations of this protocol, we envision performing calibration tasks both at the beginning and end of the self-report labelling allowing us to investigate how calibration may drift within and between sessions.

How must models of dynamic emotion evolve? Longitudinal studies will reveal how to create personalized models that evolve with the individual. Mood, life and situational context influence perception of emotional events [124] but also change dramatically over time: we wonder how repeat data collection over the course of months impacts emotion models.

How to capture a range of emotion experiences? We selected a single-dimensional scale to simplify annotation; real-life events may trigger far more complex emotion landscapes where emotions are in conflict simultaneously (e.g., feeling excited and sad about graduation). How can we make it more intuitive to document multiple simultaneous scales?

Choose or Fuse: Is report divergence an opportunity? Diverse self-reports may capture perspectives that are authentic in different ways. We have inspected characteristics of emotion self-report in the time- and frequency- domains.

Based on analysis insights, we might *choose* one approach, for its sensitivity or practicality. Or, we might *fuse* them, e.g., using discrepant moments as a spotlight on emotional conflict or low-confidence labels. We plan to develop concrete choose-fuse strategies based on focused attribute study, which also lessen intrusion on emotion experience.

6.6 Conclusion

We proposed a multipass data collection protocol to develop emotion models for real-time responsiveness in emotionally dynamic experiences. The protocol entails four sequential participant tasks: (1) emotion elicitation; (2) personal emotion calibration; and during video review, a (3) detailed interview and (4) continuous annotation of the emotion task. Using 16 participants' data, we determine that this multi-pass labeling implementation adds **versatility** to collection options, provides personalized and triangulated **insight into nuanced meanings**, and offers new options for **signal selection or integration**. We show how **emotion dynamics measures and metaphors can add value**, in particular *emotions-as-positions* or *-as-angles*; and propose promising next steps.

Ethical Impact Statement

We have proposed a novel multipass protocol for capturing and modeling high-resolution emotion experience at real-time scales. It is a personalization technique intended to benefit end-users: an automatable model evolution based on user input. While there is always potential for mal-use, this is mitigated by fundamental grounding in the individual rather than a generalized understanding of many. The investing user is the only beneficiary of model improvement; their data is of low value to others and less likely to invite exploitation.

Chapter 7

Dynamic Emotion Modelling on Incidental Emotion via Videogame Play Controls

Summary

In-body lived emotional experiences can be complex, with time-varying and dissonant emotions evolving simultaneously; devices responding in real-time to estimate personal human emotion should evolve accordingly. Models assuming generalized emotions exist as discrete states fail to operationalize valuable information inherent in the dynamic and individualistic nature of human emotions. Our multi-resolution emotion self-reporting procedure allows the construction of emotion labels along the Stressed-Relaxed scale, differentiating not only what the emotions are, but how they are transitioning – e.g., “hopeful but getting stressed” vs. “hopeful and starting to relax”. We trained participant-dependent hierarchical models of contextualized individual experience to compare emotion classification by modality (brain activity and keypress force from a physical keyboard), then benchmarked classification performance at $F1$ -scores=[0.44, 0.82] (chance $F1 = 0.22$, $\sigma = 0.01$) and examined high-performing features. Notably, when classifying emotion evolution in the context of an experience that realistically varies in stress, pressure-based features from keypress force proved to be the more informative modality, and more convenient when considering intrusiveness and ease of collection and

processing. Finally, we present our [FEEL \(Force, EEG and Emotion-Labelled\) dataset](#), a collection of brain activity and keypress force data, labelled with self-reported emotion collected during tense videogame play (N=16) and open-sourced for community exploration.

7.1 Introduction

If emotionally reactive machines could interpret the transitional nature or direction of their inherently emotional human users, responses could be designed to be contextually appropriate. Due to variations in human emotion expression and personal preferences of a desired response, such machines will likely need to be customized and tuned to the individual. In particular, a system must be able to recognize user-specific emotion *transition* through some identifiable parameter, such as intensity or polarity. For instance, when a custom emotion-aware game system estimates a user’s “anxiety” levels as low, it could ramp intensity up to a personal “frustration” threshold, to avoid game burnout.

Natural (unmediated) interpersonal emotion communication relies on many nonverbal cues: we interpret emotion expressions from others through eye contact, vocal inflections, body language and touch behaviour [7]. Using machines to recognize social touch unlocks the significant emotional content encoded in physical contact [114, 273, 321].

To model spontaneously evolving emotion in the vicinity of a participant-defined Stressed-Relaxed scale, we collected participant biosignal data while they played Playdead’s *Inside* [222], an emotionally evocative videogame. We followed the multipass data labelling protocol described in [46], recording brain activity using electroencephalography (EEG) and keypress force via a Force Sensitive Resistor (FSR)-embedded keyboard. Both have been shown to encode emotion [3, 112, 188] and are reasonable to collect during videogame play. While we considered other well-studied emotion-encoding biosignals (namely electrodermal activity, pulse oximetry, and electrocardiography), sensors that were worn on fingers or otherwise generated electrical interference with the sensitive EEG system proved unsuitable for this study.

In this paper, we present our FEEL dataset (collected under a separately peer-

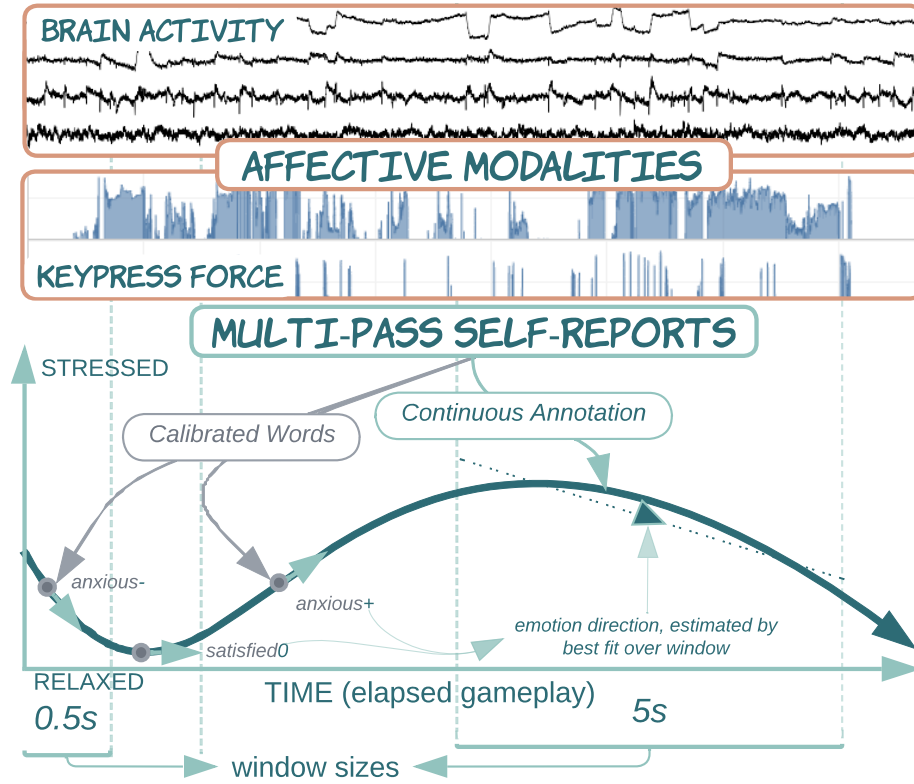


Figure 7.1: An emotion experience trajectory estimated by emotion transition. We built models on two modalities: brain activity (EEG) and keypress force (FSR), distinguishing intensifying(+), stable(0), or resolving(-) stress, at 0.5s and 5s windows.

reviewed protocol [46]) and use it to ask: **How well can we classify emotion transitions or directions using keypress force vs. brain activity collected during an emotionally evocative video gameplay?** Specifically, we demonstrate a personalized-to-participant emotion interpretation paradigm, then assess model performance, efficacy and practicality of classifying emotions as they are in flux, by comparing two distinct implicit and highly personalized expressive modalities which play out at different timescales (brain activity and keypress force). To further inform model design, particularly with respect to modality-specific frequency characteristics, we provide an evidence-based reference scale for window size se-

lection. We contribute:

1. The FEEL dataset, collected using a multipass labelling protocol featuring co-designed scales for annotating emotion self-report on keypress force and brain activity data.
2. An empirical demonstration of personalized emotion *transition* classification that distinguishes between emotion transition labels across a Stressed-Relaxed scale (e.g., cautious +, 0, or - as “feeling cautious and getting more stressed” vs. “cautious and stable” vs. “cautious but relaxing” respectively).
3. Evidence that hierarchical classification of emotion evolution along a Stressed-Relaxed dimension using touch pressure features performs nearly twice as well as continuous brain activity.

7.2 Background

Machine interpretation of spontaneous emotion requires models built on ecologically valid emotion data. From choice of expressive modality to data labelling, we ground our data collection and modelling choices in existing literature.

Affect-Encoding Modalities: Although the biological mechanisms through which emotion modulates touch are still unclear [116], touch is a concrete, perceivable and expressive act [156] and a promising modality for both inferring and influencing emotion experiences [178]. Relative to other channels commonly used in emotion research – EEG, brain imaging, heart-rate, facial configurations, body posture, speech [271, 295] – touch can be easier to harness, less intrusive to collect, and gives the participant more immediate agency in terms of behaviour compared to biological signals.

Affective touch classification has largely been based on observation and evaluation of toucher behaviour when they are prompted to reflect on a past experience [42], or to act in an emotional context [321]. While interpersonal touch pressure has been shown to communicate currently felt affect [42, 114], investigating keypress force or pressure for evidence of emotion “leakage” in the absence of communicative intent is relatively new [112]. Using pressure-sensitive keyboards,

emotion has been classified using typing pressure with up to 93% correspondence to self report (chance 17%) [188], with [112] finding a positive correlation between stress and typing force. In a mobile gameplay environment, touch pressure on an iPhone touchscreen has also been used to classify user arousal and valence at above chance levels [95]. Now, we explore how keypress force may communicate emotional transitions between Stressed and Relaxed on pressure-sensitive keys during video gameplay.

Changes in electrical potential in brain activity or electroencephalography (EEG) [168] for emotion classification is dominated by Event-Related Potentials (ERPs). However, as ERP time windows are typically constructed within 100-750ms after an event [70, 235], the ERP fails to capture emotion *evolution*, where change occurs over the course of minutes and hours [304]. Recently, 2D differential entropy-based features capturing spatial relationships and Convolutional Neural Networks (CNNs) can classify 1s data instances over emotional experiences (positive, negative, neutral) lasting 4 minutes at an accuracy of 97.10% (chance 33%) [1].

Here, we build on machine classification of emotion transition using multiscale self-reports on brain activity and keypress force during video gameplay – a dynamic emotion experience.

Emotion Self-Report: Time-varying emotion expression can be attributed to complex neurological and physiological regulation mechanisms [100], appraisal effects [198], cognition and contextual factors [194, 211]. To simplify in-lab research, computational emotion modelling often relies on emotions being represented as a point in an emotion plane along easy-to-read scales with dimensions of arousal, valence, and/or dominance [38, 42]. While these models are convenient, in real use we need to address emotion evolution over time. However, commonly used labels on the arousal-valence circumplex model [237], PANAS [309], or SAM [30] (among others) quickly become intractable for sampling at the rate of change for emotion (ranging from a few seconds to several hours [302]).

Emotion self-report with any measurement scheme raises generalizability concerns. Our understandings of the instrument scale are highly subjective [38, 148] and influenced by life experiences and personal history [20]. Any set of ground-truth labels for self-reported emotion are likely similarly personalized: e.g., one

person’s *anger* scale may be unrecognizable by another, or even by themselves at another time. In an evolving emotion experience, recognizing a particular user’s near-future emotional expression can improve the temporal and situational appropriateness of a machine response.

Emotion Modelling with Multiple Reporting Passes: With time and reflection, emotional assessment of an experience may be dramatically different from initial evocation [198, 211]. Emotions may be most intense while directly in an experience [266, 304], but articulation can only occur after some time to assess and consider the appropriate language [298]. [76] suggests the ideal window of time for emotion-naming may be shortly after an experience, to give time for processing [266] but before memory degrades [232].

Computational emotion models often rely on a single pass of emotion that is self-reported [234, 271, 295] or observed and labelled by judges. To our knowledge, our study is the first to triangulate multiple self-report methods for more reliable observation of emotion evolution.

We demonstrate the use of our FEEL dataset for exploring classification models of incidental touch pressure as a modality that captures implicit emotion expression, comparing performance to models of the more intrusive, but more studied, brain activity signals.

7.3 Dataset Description

The FEEL collection protocol [46] was a significant investment requiring ~ 400 researcher hours: each 2-hour session required a team of 4 researchers, with 2 hrs of setup, calibration and breakdown time, plus earlier piloting. As a quality assurance measure, we reviewed protocol adherence during data collection and signal quality for all 23 participants. Given our plans to publish this dataset, we used a very high standard for data quality and consistency, setting aside a participant’s entire record where at any point during the session there was any suspicion of excessive noise in EEG data, equipment malfunction, synchronization mishap or possible recording errors. This left us with 16 publishable records (7 omitted due to any combination of the above set of minor issues).

The FEEL dataset consists of comma separated value (.csv) files organized by

participant. Video data is excluded for participant privacy. Analyses start with this 5.4GB dataset, available at https://www.cs.ubc.ca/labs/spin/FEEL_dataset [to be posted upon acceptance].

Data Capture and Preparation: As part of recruitment, participants completed a questionnaire adapted from the Trait Meta Mood Scale (TMMS) [244]. Based on these results, we invited only those scoring with high emotion clarity and low emotion suppression based on their responses.

Of the N=16 participants, 8 are female and 8 male; 8 between 19-24 and the other 8 between 25-34 years of age. All played videogames regularly from a few hours a month up to 4 hours daily, nearly all of whom report 1-6+ hours per week; none had played *Inside*. All were compensated \$30 for the 2-hr data collection session.

Data collection was conducted in four steps [46]:

1. *Initial Gameplay* generated streams of participant brain activity (EEG) and keypress force from an FSR-embedded keyboard timestamped from the first keystroke, indicating the start of gameplay.
2. In *Word Scale Calibration*, participants placed pre-selected emotion words relative to one another on a Stressed-Relaxed emotion scale.
3. In the first self-report cycle (*Calibrated Interview*), participants then reviewed and annotated the gameplay video with their calibrated word sets.
4. Finally, in the second self-report cycle (*Continuous Annotation*) they used a 1D joystick (position sampled at 256Hz) to annotate the video.

We timestamped data streams with corresponding frames from the Initial Gameplay video, where participant gameplay averaged 13:24 minutes (min 8:25, max 21:37, SD 3:53).

Brain Activity Data Stream (EEG): Participants were instructed to minimize conscious movement; researchers noted sessions with excessive motion to check for unusable EEG data (deciding to omit it if so).

We captured brain activity data using EGI's EEG 400 system¹, sampled at 1kHz, with a band pass filter of 1-50Hz applied in post-processing. We followed

standard practice in removing high frequency jitter and 60Hz mains noise [326] while retaining α , β , θ , and γ frequency bands (associated with emotion processing [2]). We did not downsample because (a) we were able to efficiently capture important dynamics using spectral-domain features and (b) we are still exploring which frequency components are important.

We checked classification performance over a number of data cleaning procedures using MNE-Python tools², including artifact removal and baseline correction by the entire gameplay duration, and by adjacent windows. We also tried applying Independent Component Analysis (ICA) to address eye blinks and removing channel segments with exceptionally high noise levels. These procedures yielded no notable classification improvement or a marginal performance decline over 30 training and testing iterations. So, we report results and publish the dataset with minimal pre-processing³, largely leaving EEG data “alone” as recommended by [69]. We used this data version in the classification models reported in this paper.

Keypress Force Data Stream (KFP): We embedded force-sensitive resistors (FSRs) on game-specific control keys (four direction keys and ALT) on a standard keyboard. Force ranged from 0 (no contact) to 1023 units ($\sim 1\text{kg}$)⁴. We downsampled FSR data from 52Hz to match videogame framerate at 30Hz.

Timeline with Calibrated Words (TwCW): The Timeline was created from collection sequence Steps 2 and 3.

Word Calibration: Following gameplay, players calibrated a Stressed-Relaxed emotion scale, contextualizing scale-points with memories of their recent gameplay experience and marking 13 pre-selected emotionally “Calibrated Words”: *Cautious, Satisfied, Hopeful, Frustrated, Anxious, Nervous, Threatened, Resigned, Alert, Accomplished, Fearful, Dread, and Curious*. Participants were also allowed

¹EGI EEG system details: <https://www.egi.com/research-division/eeg-systems/geodesic-eeg-systems>. Model 400 features a 64-channel Routine Hydrocel geodesic sensor net, proprietary NetStation data collection and visualization software.

²MNE tutorials available at <https://mne.tools/stable/index.html>

³Included processing ensures labelling format consistency and time alignment across data streams. The FEEL dataset is published unfiltered with no artifact, segment, nor baseline correction. Any processing prior to classification is described in Section 7.4 - Methods.

⁴As defined by the FSR specifications available commercially at <https://www.robotshop.com/en/force-sensing-resistor-fsr.html>.

to write-in up to two additional words. This individualized calibration step contextualizes how each person perceives and uses these words with respect to the Stressed/Relaxed dimension, improving participant-researcher grounding on language usage [38, 46].

TwCW Construction: Players reviewed their gameplay video, annotating (calibrated) emotion words at timepoints associated with strong emotion ($\mu=0.05\text{Hz}$, $\sigma=0.015\text{Hz}$). To construct the TwCW, we associated each interview annotation with the calibration value for that word, at the annotated gameplay timestamp.

Continuous Annotation Stream (CA): In the second gameplay review, the CA is generated from a non-biased joystick (holds last position rather than returning to centre) tracing an emotion time series, where the resulting curve is a proxy for a participant’s true emotion trajectory between Relaxed and Stressed over the timeline of the gameplay experience. Joystick position readings were matched with video frame rate of 30Hz to ensure alignment with video playback. We smoothed analog jitter in the joystick data with a simple moving average filter, then normalized range to [0:1].

Figure 7.1 highlights the data collected during the study: player-specific gameplay streams (EEG and FSR), emotion word calibrations, and the TwCW and CA – two time-series of emotion self-report annotated on the same dimensional plane of the Stressed-Relaxed scale over the gameplay timeline.

7.4 Methods

To demonstrate personalized emotion transition classification using our FEEL dataset, we created participant-specific hierarchical multi-label models to leverage several benefits, in the context of multi-label classification tasks featuring multiple label streams – here, emotion words and quantitative stress measures. By incorporating a hierarchical structure into the model, we capture the complex and dependent relationships that may exist between labels, thereby improving classification accuracy [272, 297]. This approach is also more flexible in handling different types of label streams, and comprehensive in its view of the individual’s emotional state in both brain activity and keypress force.

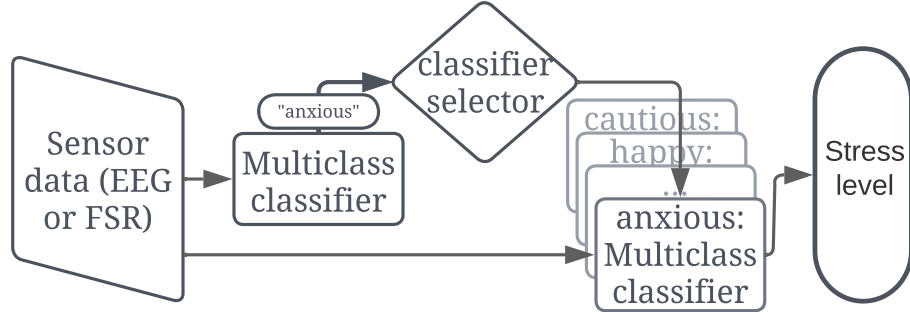


Figure 7.2: Our hierarchical machine learning framework employs a two-tiered classification strategy. Initially, a local multi-class classifier is deployed at the parent node level to identify the primary category, termed as the "Calibrated Word." Subsequently, for each emotion word identified in the first step, dedicated models are trained. These models are designed to predict one of three potential outcomes related to the "Stressed" category. This architecture allows for a nuanced understanding of the data by first broadly categorizing the input and then applying specialized models for a detailed analysis within each category.

7.4.1 Data Instances: Labels and Window Lengths

We aligned FSR, EEG, and emotion self-report time series, dividing streams into non-overlapping, equal-duration windows. We analyzed window lengths of 0.5, 1, 2 and 5s, spanning ERP window range [70, 235] up to perceived emotion duration of "a few seconds" [303, 304]; 1s and 2s windows match other emotion-related classification studies [42, 182, 308]. Results from intermediate lengths followed the trend set by the extreme values, so we report only 0.5 and 5s for brevity.

A single data instance consists of features and labelled emotion class calculated from data within one window. Across all participant sessions, we collected an average (over all participants) of 1435.13 data instances for 0.5s windows ($\sigma=405.51$) and 142.63 instances for 5s windows ($\sigma=40.75$).

To implement hierarchical multi-label outputs, we used the Python package Hi-Class [196] with algorithms implemented in scikit-learn [219], XGBoost [52], and the Pytorch framework [218]. In a 2-stage approach, we first trained participant-specific multi-class models that output the emotion words from the TwCW, then trained a classifier from the CA by each Calibrated Word, outputting binned direc-

tion values (slope of best-fit-line as in [46]). The same algorithm was used for both stages by modality: classical ML for FSR and CNNs for EEG.

The resulting label set across 16 participants consisted of approximately 11 ($\mu = 10.94$, $\sigma = 1.91$) distinct calibrated emotion *word* labels, out of a possible 15 calibrated emotion words (13 provided and 2 write-ins per participant, Table 7.1).

For each word, there are three possibilities regarding transition direction; e.g., *Nervous* could be *nervous+*, *nervous0*, and *nervous-*, representing being nervous but with intensification along the *Stressed* scale, stable stress, and resolving stress respectively. When looking at transition directions for each word used by each participant, we found that all three possibilities appear for most words, except in cases where an emotion word is mentioned only once or twice (such that it could not be associated with three distinct directions). Observed distributions were $[\mu, \sigma]$: $[2.76, 0.53]_{5s}$ and $[2.96, 0.22]_{0.5s}$.

Figure 7.2 exemplifies the hierarchical process with two streams of self-reported emotion. Where window boundaries do not coincide with a logged data point, we imputed with the previous data point, turning our time-series into a higher-resolution stepped signal. We resolved windows containing multiple labels by using mode for the Calibrated Words and the slope of the best fit line in the continuous annotation.

Table 7.1: Full list of Calibrated Words used by at least one Participant in their TwCW.

Calibrated Word	Number of Participants	Calibrated Word	Number of Participants
Anxious	15	Confused*	11
Frustrated	15	Curious	11
Dread	14	Resigned	10
Indifferent*	14	Threatened	8
Satisfied	14	Annoyance*	5
Hopeful	13	Resolve*	4
Accomplished	12	Excited*	3
Alert	12	Clueless*	1
Cautious	12	Triumph*	1

Participants used 11 of the 13 provided words (none spoke of feeling *Fearful* nor *Nervous* during the interview stage so both are omitted). Starred * words are participant-generated write-ins.

7.4.2 Force Sensitive Resistor (FSR) Data:

FEEL’s keypress force (FSR) data exhibited an average of <1 distinct keystrokes per window, contraindicating deep learning models. We extracted features from keystroke activity, frequency, and statistical analysis, generated data instances aligned with brain activity (0.5s and 5s), and performed model selection with classical machine learning models.

Data Preparation: To mitigate FSR signal noise while maintaining the overall shape of a keystroke, we applied an Exponentially Weighted Moving Average (EWMA) [134] with smoothing factor $\alpha = 0.5$. We aggregated game keypress activity from the original game-control keys (denoted A0-A4 in the dataset) into two additional channels as ‘**composite keys**’, computing over all keys the force sum (A5) and maximum (A6), resulting in a total of 7 keypress channels.

Frequency and statistical features: Based on previous studies of emotion expression of social touch pressure [42], we calculated a set of descriptive statistics for each window of pressure data – minimum, maximum, variance, mean, area under the curve, and sum of absolute differences. From the same windows, we calculated the most prominent frequency (amplitude and frequency bin), amplitude variance, amplitude mean, and peak count for frequency-domain features [42].

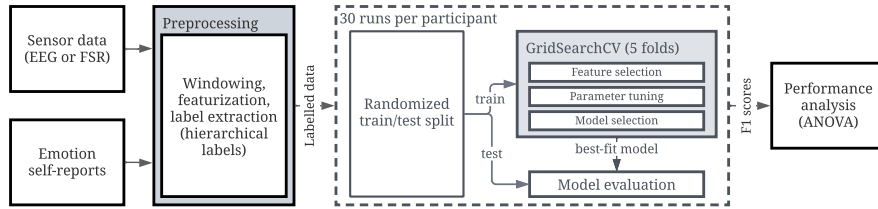


Figure 7.3: Pipeline for model selection and evaluation. We performed grid search CV ($k = 5$) on the training set to tune hyperparameters and select best-fit models for FSR data. The models were then evaluated on an unseen test set to calculate performance metrics. We repeated this process 30 times per participant, and report mean test scores across the 30 runs and 16 participants.

Keystroke features: Since participants activated keys based on gameplay rather than typing, certain features of keystroke dynamics – such as travel time between

keys – are less relevant here. We therefore calculated touch features highlighting fluctuations in force and duration in both time and frequency domains [42, 151]. We also borrowed parameters related to the Attack Decay Sustain Release (ADSR) envelope [140], commonly employed in synthesizers to describe piano keyboard output. For each keystroke in a window, we calculated: keystroke duration (in ms), peak count, amplitude of maximum peak, time from keystroke start to maximum peak, time from maximum peak to key release, force variance, average force, and area under the keypress curve. Parameters are aggregated by taking the mean over each data window.

For the purposes of multi-modal window alignment and the simulation of real-time application of emotion classification on keypress force, we used uniform data windowing. However, we note that distortion may occur where keystrokes cross window boundaries.

7.4.3 EEG Data

We calculated Differential Entropy (DE) for the 5 frequency bands demonstrating activity during emotion expression [1, 75]: δ (1-4Hz), θ (4-7Hz), α (8-12Hz), β (12-30Hz) and γ (30-50Hz). For each band, we calculated the difference between channel pairs to create a 2D Asymmetrical Map (AsMap) feature [1]. The resulting feature is an image with size 64×64 and a depth of 5 frequency bands.

7.4.4 Classification Model Implementation

To compare EEG- vs. FSR-based models classifying emotion transition, we ran 30 iterations of 5-fold cross-validation (training and validation sets randomized every iteration). Figure 7.3 summarizes the overall experimental pipeline.

FSR: We performed grid search cross-validation (CV) ($k = 5$) to select the best-fit model by participant among seven machine learning models. Due to the sparseness of the FSR data (low sampling rate with some keys pressed in only a few brief instances), we elect to compare performance across Extra Trees, Random Forest, AdaBoost, Gradient Boosting, XGBoost, Logistic Regression, and SVM [52, 150, 219], options that are more amenable to the size and scale of this data than deep learning models. Given the high dimensionality of our feature set ($d = 82$ features

per participant), we selected features by employing a zero variance threshold to remove all constant-valued features and use recursive feature elimination (RFECV) [106] with CV ($k = 5$). We report mean test scores over the 16 participants after 30 runs using the best-fit model for each, with a 70/30 training-test split ratio.

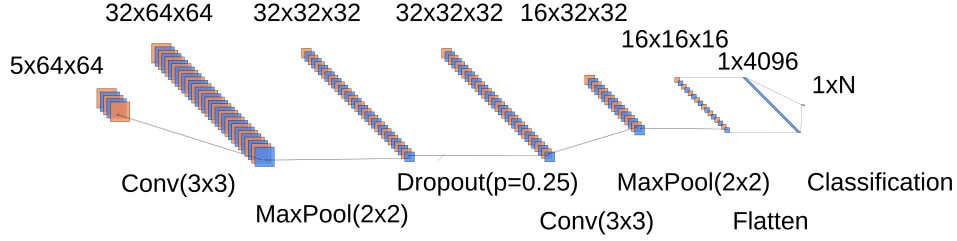


Figure 7.4: Structure of the EEG CNN model for classification where each convolution layer uses a 3×3 kernel (of depth 32 and 16 respectively) followed by a ReLU activation function. The inputs to the model are the $5 \times 64 \times 64$ AsMap features [1], while the output is the class output ($N = 3$).

EEG: We used a CNN model with a 2D feature set to take advantage of the automated learning demonstrated by deep-learning models. In the interests of balancing model complexity with overfit risk [51], we implemented the structure proposed by Ahmed *et al*(2022) [1] – a 2-layer CNN using 3×3 kernels and 2 Max Pooling layers – for affect classification, adjusting the input size to $5 \times 64 \times 64$ to account for the size of our features. We created train and test sets with the same 70/30 split ratio as with the FSR data. Figure 7.4 summarizes the CNN architecture.

We performed grid search CV ($k = 5$) on the train set to tune the number of epochs (5, 10, 20), the batch size (128, 256, 512) and the learning rate (10^{-3} , 10^{-4} , 10^{-5}) to select the best participant-specific hyperparameters for our model. Larger epoch sizes (≥ 100) were omitted from the search space since similar training performances were observed, while being resource intensive. Once we obtained the parameters that maximized the macro-hierarchical F1-scores (as defined by Miranda et al. [196]), we trained the participant-specific model 30 times on the full training set, each time using the unseen test set to calculate performance metrics. We report mean test scores from 30 runs.

7.5 Classification Performance by Modality

We analyzed macro hierarchical F1-scores [196] by model and window size (Table 7.2) finding that classification performance monotonically increases with window size. For brevity, we report in depth on 0.5s and 5s windows. With two modalities and two window sizes, our data does not pass Levene’s test for equality of variances ($F(3,1916)=51.0$, $p < 0.001$), so we report results using a two-way aligned rank transform analysis of variance (ART ANOVA), implemented with R’s AR-Tool [316]. All reported effects are statistically significant at $p \leq 0.001$. The main effects of affective modality (M) and window size (W), and interaction effect (W/M) yield F ratios of $F_M(1, 1916) = 5283.98$ ($\eta_p^2 = 0.733$), $F_W(1, 1916) = 56.88$ ($\eta_p^2 = 0.028$), and $F_{W/M}(1, 1916) = 285.26$ ($\eta_p^2 = 0.130$).

Table 7.2: Hierarchical classification scores for each (W)indow / (M)odality where the best combination is **5s-FSR**. All W/M models exceed chance by ~ 2 -4x.

W/M	F1-Score	Precision	Recall
5s EEG	0.415 ± 0.110	0.415 ± 0.109	0.422 ± 0.123
0.5s EEG	0.494 ± 0.070	0.544 ± 0.118	0.469 ± 0.090
0.5s FSR	0.686 ± 0.039	0.681 ± 0.036	0.682 ± 0.037
5s FSR	0.823 ± 0.012	0.827 ± 0.013	0.825 ± 0.013
0.5s chance	0.215 ± 0.010	0.215 ± 0.010	0.215 ± 0.010
5s chance	0.216 ± 0.009	0.216 ± 0.009	0.216 ± 0.009

Scores are calculated over 480 hierarchical metrics (16 participants \times 30 runs, average macro hierarchical F1 taken over all classes).

We ran post-hoc tests using a Holm correction to further investigate the individual mean differences in Table 7.2 (significance at $p_{Holm} \leq 0.001$ unless indicated). Results show that (1) mean F1-score was significantly greater for FSR-based models than EEG-based models; (2) mean F1-score increased with window size, with 5s windows performing strongest across modalities; and (3) FSR at 5s windows performed best overall. Additionally, we found that the chosen CNN parameters for batch size and epochs tend to differ by participant, while the learning rate remained stable. For 0.5s, the optimal parameters by participant were seen for batch sizes of $\mu=160.0$, $\sigma=96.0$ and training epochs of $\mu=13.8$, $\sigma=6.5$; for 5s, $\mu=248.0$, $\sigma=139.0$ batch size and $\mu=11.6$, $\sigma=6.1$ epochs. In all cases, loss curves stabilize by 15 epochs, suggesting diminishing returns in classification perform-

ance with additional training epochs.

7.6 FSR Feature Analysis

For insight on how features inform classification, we ran RFECV on the feature set of both FSR models (0.5s and 5s windows), and grouped selected features by type – pressure-based (direct measures of keypress force), time-based (measures of duration), and frequency-based (FFT-based features). We analyzed model performance using F1-score for feature group. Figure 7.5 summarizes the top performing feature groups.

Our data for both models (0.5s and 5s) again does not pass Levene’s test for equality of variances ($F_{0.5s}(2,39356)=831.74$, $p_{0.5s} < 0.001$; $F_{5s}(2,39356)=906.32$, $p_{5s} < 0.001$) with three feature groups, so we report F1-scores after two one-way ART ANOVA for each window size. Main effects of both tests are statistically significant at $p \leq 0.001$ significance, yielding F ratios of $F_{0.5s}(2, 39356) = 1049.3$ ($\eta_p^2 = 0.051$) and $F_{5s}(2, 39356) = 761.97$ ($\eta_p^2 = 0.037$), respectively.

To investigate individual mean differences, we ran post-hoc tests using a Holm correction significance at $p_{Holm} < 0.001$ unless otherwise indicated. The mean F1-score was significantly greater for models that rely on **pressure-based features, for both window sizes**, followed by time-based features.

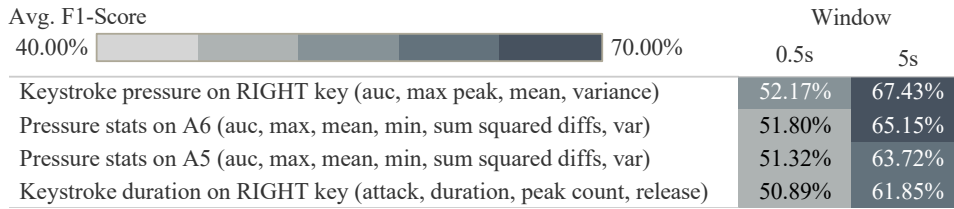


Figure 7.5: Relative feature performance by window size. Darker cells indicate frequent selection of better-performing features. The RIGHT directional key is used to advance the character – and game storyline – through the side-scrolling game. A5 corresponds to the sum of the pressure across all keys, while A6 corresponds to the max force over all keys.

7.7 Discussion and Future Work

Here, we reflect on our research question, and how our findings can inform the use of touch pressure data in modelling dynamic emotions and contribute to the development of emotionally responsive devices.

7.7.1 Real-Time Predictors of Dynamic Emotion

Longer Time Windows Favour Keypress Force: For personalized classification models of evolving Stress built on participants screened for high emotion clarity, FSR models perform better than those built on continuous EEG for both window sizes we analyzed. Individualistic emotion evolution inherent in real life events, particularly when reflecting or reacting to memory retrieval, may require more than 0.5s [279, 302]. We posit that longer windows will better capture lower-frequency information and thus benefit manual keyboard interactions for models of keypress force, but may blur the picture of higher-frequency brain activity features [308, 326].

Manual touch pressure encodes valuable emotion content: Our feature extraction techniques were informed by analyses of a variety of affective touch interactions: keystroke dynamics in typing behaviour [188], pressure and location features from social touch [42], and ADSR features from sounds produced from a music keyboard [140]. Feature evaluation reveals that of the 20 most important features from all three domains, 16 are pressure- or force-related. Increases in typing force were previously known to correlate with higher stress experiences [112], and machine-mediated social touch [151] has been differentiated by variations in pressure. Now, we have evidence that Stressed-scale emotion expression can also be captured implicitly through keypress force using an easily modified videogame keyboard. We continue to investigate other contexts and emotion scales where we subconsciously express emotion via touch pressure, leaving dimensional examination of dynamic emotion evolution and touch pattern correlates to future work. In the meantime, we posit that the information available by tracking pressure in devices where interactions feature manual affective touch outweigh the cost of adding this functionality.

The case for emotion transitions – timing matters: When modelling human emotions, we may consider how the emotion space changes over time: when we feel sad, it may be easier to get angry than calm, despite these emotions being separated by comparable Euclidean distances on the Affect Grid [237]. An emotion experience can feel more like a trajectory over a constantly changing landscape than a point [38]. After studying the evolution of *stress*, we infer that predicting direction of an emotion trajectory may be particularly important when delivering interventions for emotion regulation. For example, strategies may differ for the *onset* of anger vs. after rage has *cooled* [266].

7.7.2 Building Effective Models for Dynamic Emotion Prediction

Potential confounds: First, we point out that there are a number of potentially confounding factors, including (but not limited to): participant interest in, and proclivity for, this video game genre; fluctuations in skin conductivity; extraneous motion; model complexity vs. availability of training examples per class; as well as, cognition in action planning; personal experiences of Stress-Relative emotions; individual differences in the ability to express, appraise, and resolve emotions. We minimized these limitations through participant screening, personalized word calibration, multipass data labelling for richer experience capture, and individualized emotion classification models. However, they may still have influenced the reported classification performance.

Modality capture: The collection of this dataset was time-intensive and effortful, in large part due to setup and calibration of the EEG data collection system. Given EEG signal sensitivity to surrounding conditions as well as collection effort and intrusiveness, the comparable-to-better classification performance of FSR signals for emotions adjacent to the Stressed-Relaxed scale means that under certain conditions – e.g., slower evolution as for Stress, emotion reflection tasks requiring appraisal or memory retrieval [279], low compute and/or time resources, or prioritization of personalized over general models – we are hereby able to recommend reliance on, or the addition of, keypress force or other manual touch data for emotion interaction.

Labelling effort: Collecting multipass emotion self-reports affords rich triangula-

tion of a numerical emotion rating onto personalized emotion scales. But it also incurs a time cost: altogether, personalized calibration, emotion elicitation, interview, and continuous annotation take 3 to 4 times as long as the emotion elicitation task alone. Where tasks run long, multipass reviewing procedures require careful consideration to ensure annotation can occur contemporaneously without interfering with the natural evolution of the emotional experience.

Emotion elicitation and affect scale: Calibrating how users placed emotion words on a Relaxed-Stressed scale allowed us to simultaneously pool data and personalize models. While participants had personalized understandings of the measurement scale, they all engaged in the same emotion elicitation experience (a horror video game).

For personalized models to work “in the wild”, they must be built on participant-defined emotion experiences that evolve longitudinally and spontaneously. Human emotional experience is ever-evolving; so also must be the calibrated scales, training data, and accompanying models across multiple named emotions and touch interaction patterns. Future work examines how longitudinal calibration can trace evolution of emotion models over multiple data collection sessions.

Context Matters in Personalized Emotion Models: A deployed model could face a wide range of priorities. Naturalness of a responsive agent may value minimal latency over accuracy. In other situations, some scenarios may be more important to capture accurately (‘something’s wrong’) than others (‘everything’s fine’). Machine learning accuracy metrics are useful for comparing performance, but for contextually effective machine responses, new metrics may be necessary to reflect the nuances of the overall experience.

7.8 Conclusion

We present the FEEL dataset, the first of its kind: affective multimodal data (brain activity and keypress force estimated by EEG and FSR) collected during an emotional videogame experience and labelled using a multipass emotion self-report described by [46] – resulting in multi-timescale, and personally calibrated emotion labels rooted on the Stressed-Relaxed scale. This paper describes the dataset and the specifics of its collection, and demonstrates participant-dependent machine

learning classification performance differentiating emotions in **transition** – e.g., whether one’s *stress* is growing or resolving, benchmarked here at $F1 = 0.82$ at the best case (chance $F1 = 0.22$, $\sigma = 0.01$). We invite the community to explore other computational strategies and advance the exploration into dynamic emotion classification.

Comparing classification performance over factors of window size, feature set, and modality, we find that, overall:

1. Window sizes influence recognition behaviour for both brain activity and touch pressure, the choice of which depends on intended observation (longer windows are better able to capture slower changes but shorter windows can capture high frequency activity)
2. Feature evaluation of the FSR feature set reveals that pressure features used in machine-mediated social touch rank highest in terms of selection frequency.

From these findings, we propose that emotion interaction systems should (1) consider window size in labelling; and (2) improve emotion recognition opportunities by incorporating pressure sensors where manual human touch is enacted.

Chapter 8

Collecting and Labelling Training Data for Dynamic Emotion Classification: A Proof-of-Concept

Summary

Our emotional expressions and experiences are highly contextualized to personal histories and may not be easily generalizable across populations. To open up new possibilities for enhancing emotional experiences through haptic feedback, and foreshadowing the development of personalized and immersive technologies that capture and respond in time with user emotion, we present a proof of concept for system training. Intended for personalized models of authentic emotion, this approach features true emotion expression evolving from expressive, naturalistic touch elicited through user storytelling. While participants recalled a personal story that evokes strong emotions, such as describing memories around the best period in their life, we collected touch data using a 10 inch x 10 inch touch sensor embedded on a soft cushion, along with heart rate and Galvanic Skin Response (GSR) measurements. Through multiple sessions with $N_{participants} = 5$ providing between

1 to 3 sessions each, we gathered $N_{sessions} = 10$.

From participant-reported continuous annotation on their emotion evolution curve (drawn with a non-biased joystick that stays in place when moved), we define “emotion direction” labels as calculated by the angle of inclination of the slope of the curve. Intended for customized device applications, we demonstrate classification accuracies of emotion direction at 58% (SD 18%), exceeding chance at 25% using our procedure for participant-dependent models. Results show that negative-valenced prompts generated emotional stories nearly 2x longer than positive-valenced prompts, joystick-labelling of emotion evolution may have multiple interpretations or strategies, and that stories about unresolved emotions around an event evoke the most authentic and intense emotions, no matter how long ago the original event occurred.

8.1 Introduction

For emotionally responsive devices to track and recognize our complex and rapidly evolving emotion experiences, they must be trained with data from representative dynamic human emotion expression. However, collecting such training data that is both labelled at the timescale of emotion evolution and also reflects authentic and spontaneous naturalistic expression is widely acknowledged as a challenge for ‘in-the-wild’ emotion recognition [73, 174] for all technical sensing modalities.

when allowed natural expression, emotions generally arise without our cognitive assessment or naming of that emotion – suggesting that the labelling or reflection process could interfere with or alter naturalistic emotion onset [211]. Therefore, reflection and labelling can only be done after expression but not so long after that we lose the fine resolution and clarity of the memory. One approach to balancing the needs for unencumbered experience with good memory access is to first fully record the emotive experience with cognitively unintrusive mechanisms (e.g., tracing the arousal-valence grid with a cursor [61]), then elicit self-reports in post-review. Common signals examined for emotion classification, which are available for experiential collection with relatively low emotional interference, include brain activity [155], skin conductance, heart and respiratory rate and variability, eye gaze [139] and touch behaviour [114] – all known to embed or

reflect emotion content.

We focus on touch as a crucial avenue for interacting with both our environment and the people around us. Understanding the relationship between naturalistic touch and authentic emotion is crucial for gaining insights into human emotions [114] and developing touch-sensitive devices capable of being emotionally responsive to in-the-wild dynamic emotion in realtime. Additionally, the highly individualized nature of touch expression makes for strong identification of the individual user but presents a challenge for generalized, or participant-independent, emotion classification [42, 87]. For touch-sensitive emotionally-responsive devices that are intended for use with a small set of users (perhaps within a household), personalizing the underlying classification model could greatly improve usability (much like personalized touchscreen keyboards that adapt to user-specific typing behaviour [85]).

As a proof-of concept of this approach, in this work we explore the feasibility and ease of use for building an emotion-labelled training set of natural and spontaneous touch occurring during an emotionally prominent experience. The protocol builds on the multi-stage data collection procedure described in Chapter 6 (published as [43]) using a personalized storytelling task for emotion elicitation. Just prior to the storytelling session, participants were given a soft cushion to hold with instructions to place their hands on it as we set up the equipment. The cushion was wrapped in a custom fabric touch sensor and by leaving the hands free from any other specific activity, we could record their natural and spontaneous touch expressions arising from emotionally charged storytelling as they played out on the cushion. Participants also wore physiological sensors that recorded skin conductance and heartrate; these are known emotion encoding signals [77, 104, 155] collected here to check for emotional corroboration with touch expression. Following the emotion elicitation task, participants engaged in a multi-stage, multi-resolution emotion self-report process very similar to [43].

Approach: Our objective was to figure out a way to achieve both emotional ecological validity and computational rigor. To this end, we developed a machine-learning modeling pipeline on a relatively small sample of training data collected in this manner. In an exploratory approach, we assessed performance at each stage

of model-building, and iterated on label binning, label distribution, and classification paradigm as we proceeded.

After piloting our procedure on two individuals, we collected data on five more individuals (mean = 29 years old; sd = 6.6 years). Two people returned for three sessions each, one for two sessions, and the other two contributed a single session each, for a total of ten sessions.

This chapter focuses on how we constructed the training dataset and analyzed the participant responses, to contribute:

1. a proposed procedure for building a more ecologically valid training dataset for personalized emotion recognition;
2. a software system designed and confirmed to facilitate this complex multimodal data collection and multi-step labeling protocol; and
3. reflections and recommendations on the full process pipeline, leading to improved future iterations of both collection and realtime classification protocol.

For a future larger study, we can then confidently deploy a protocol based on findings here.

The remainder of this chapter will feature a brief background on emotion labels for machine interpretation and associated protocols (Section 8.2), a description of the data collection and labelling procedure (Section 8.3), and an analysis of the participant experience (Section 8.4). In addition, for context we summarize the outcome of the full study in Section 8.5 (full details were the focus of co-author Guerra and will be reported separately). We finish this chapter with reflections.

8.2 Background

Our approach is rooted in advancing touch-centric emotional interaction between a human and a responsive device, such as a robot. Here, we examine relevant work in computational models for emotion recognition and the data that drives them.

8.2.1 Ecological Validity of Emotion in HRI

Human-machine interactions can be socially and emotionally fraught even when this is not intended [148]. For example, a study on robot touch interpretation [56] shows that when a non-social robot “violates” an expected handover event (doesn’t successfully hand over a block to a human collaborator), the follow-up robot-initiated tap on the arm is perceived as an apology or other attempt at social repair. This suggests that while people report wanting to interact with machine agents as they do with people [88], it may be more a socio-emotional reflex than a considered preference. Thus when we design for human interaction with machine agents, we must acknowledge the emotional impact [148] inherent in the narrative users build as they assume human social touch conventions and the accompanying emotion expressivity. This is particularly true when we leave the confines of the lab environment and deploy robots into socially defined spaces like workplaces, care environments, or homes [147].

8.2.2 Protocol for Eliciting and Labelling Dynamic Emotion

Designing for ecologically valid Human-Robot Interaction (HRI) means that we have to take into account how human emotion actually progresses. Classifying emotion as *state* has many practical benefits for machine recognition – not least of which are the popularity of reporting scales used for identification and measurement. Instruments like Russell’s Circumplex Grid [237] and the Self Assessment Manikin or SAM Scale [30] distinguish emotions into two or three orthogonal dimensions of arousal, valence, and dominance. By offering a forced choice of simple and straightforward classes for data labelling, they assume a (albeit computationally convenient) model of a single time-invariable or static emotion per rating. However, emotions rarely fit into convenient boxes, rather they are complex and dynamic in situation-dependent ways. Who we are with, how recently our physical and emotional needs have been met, and why we are here now with all the baggage of our cultural and personal history [20] influence how we feel and how these feelings will evolve throughout the course of a single event or experience, as well as over longer extents of time [169]. Operationalizing concepts rooted in emotion dynamics for computational applications may require introducing new metaphors

of emotion or class labels that can capture transitional emotional experiences as they happen [39].

We use storytelling, an elicitation technique based on relived emotion, where researchers use semi-structured interviews to ask participants to retrieve memories of a significantly emotional event [57, 179, 180]. Our own previous experience with this technique has demonstrated its potential for bringing up strong emotions in participants [42]. However, for this present study, we are not asking participants for a target emotion [42] but rather for a larger story that explores many strong and evolving emotions associated with a more general (though still emotionally charged) prompt. To identify the emotions experienced through the storytelling, we use another previously explored technique: multistage labelling that incorporates emotion word calibration, an interview, and continuous annotation of emotion [43, 44]. By using the slope or angle of emotion inclination as a way of determining emotion trajectory, we can add a dynamic element to describe how an emotion may evolve.

8.2.3 Spontaneous Emotion in Training Data for Machine Classification

Machine recognition of human emotion requires training data of naturalistic and spontaneous emotion – a well-studied example is the Dataset for Emotion Analysis using Physiological Signals (DEAP) [159] representing emotions via brain activity and facial expression while participants listened to musical stimuli. To our knowledge, a dataset of natural and spontaneous emotion that includes touch interaction does not currently exist. Publicly available datasets of labeled emotion data synchronized with expressive touch data include ‘performed’ social touch where gestures are acted based on a set of instructions like the Corpus of Social Touch [151] and the Human-Animal Affective Robot Touch [41], used as affect classification challenges [152]. In a more naturalistic open-sourced example, incidental emotion is labelled on keypress force during videogame play [44]. However, this is a very context-specific environment and may not be representative of general spontaneous emotionally expressive touch.

To better examine the limitations and challenges of creating such a dataset for human-machine touch interaction, we begin by collecting multiple samples from a

small number of individuals and exploring the ecological validity of the procedure for machine classification of dynamic emotion via naturalistic touch expression.

8.2.4 Comparing Modalities

We choose to examine modalities based on the primary interaction for our intended application (touch-sensitive machines) to better understand how classification of emotional touch compares when supported by other data streams that offer known emotional encoding and can be collected with minimal interference.

Skin conductance (measured as galvanic skin response, or GSR) and heart rate variability (HRV) offer a proxy measure of emotional experiences [102, 163]. Both GSR and HRV are influenced by the autonomic nervous system which regulates the body’s internal organs and controls the body’s response to stress [163]. Sweat, particularly in hands and feet, can be a response to emotional stimulation and other psychological processes [163] and makes GSR a sensitive measure for emotional arousal. HRV has been used as a measure of emotional regulation [8] and has been shown to be associated with various emotional states, including anxiety, depression, and stress [261].

Both can be measured with small and relatively unintrusive sensors worn on the hand not otherwise engaged in naturalistic and emotionally expressive touch.

8.3 Naturalistic Data Collection Procedure

We followed a mixed-method, multi-stage emotion self-reporting procedure [46] to achieve high-resolution, temporally sensitive, and multifaceted data on emotional experiences while upholding the reliability and validity of the collected data. By acknowledging and tracking the temporal dynamics of emotions [169, 250], we can gain valuable insights into the progression and evolution of emotional experiences, advancing the move from in-lab to in-the-wild emotion recognition.

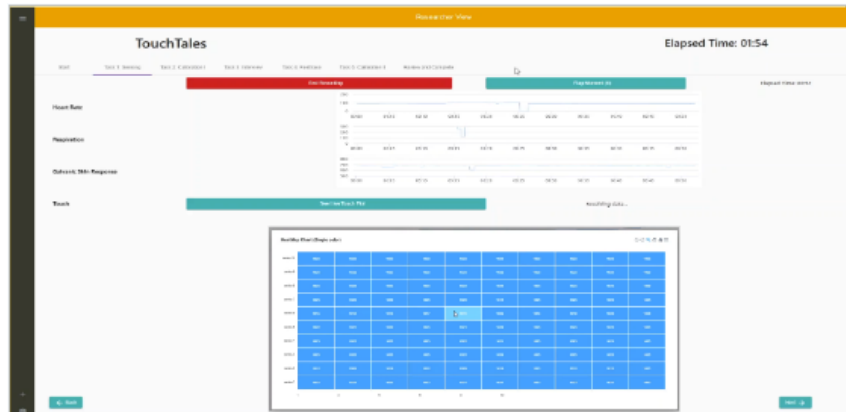
8.3.1 Software and Interface

Dynamic emotion classification requires the synchronization of all expressive modalities (here: touch, GSR and heartrate) as well as the self-report emotion labels to the same timeline [44]. We present a modular emotion dynamic annotation tool that

generates experience-aligned high-density time series of dynamic emotion on user-defined emotion scales. We built this tool to label affective touch behaviour during an emotional storytelling episode. As such, the interface features a live touch sensor heatmap, providing valuable potential for building insights on how touch behaviour evolves with emotion from one millisecond to the next – an essential first step to developing affective touch-sensitive devices that respond to dynamic emotions in realtime.

This Javascript tool is designed to facilitate emotion studies involving multiple self-report stages, with customizable modular components for specific study needs [43]. We demonstrate easy synchronization for multimodal data collection, providing support for microcontroller-driven (e.g., Arduino) live sensor synchronization and visualization, including a live heatmap for touch-sensitive interfaces, multimedia custom labeling on timestamps, and audiovisual biosignal sensor synchronization.

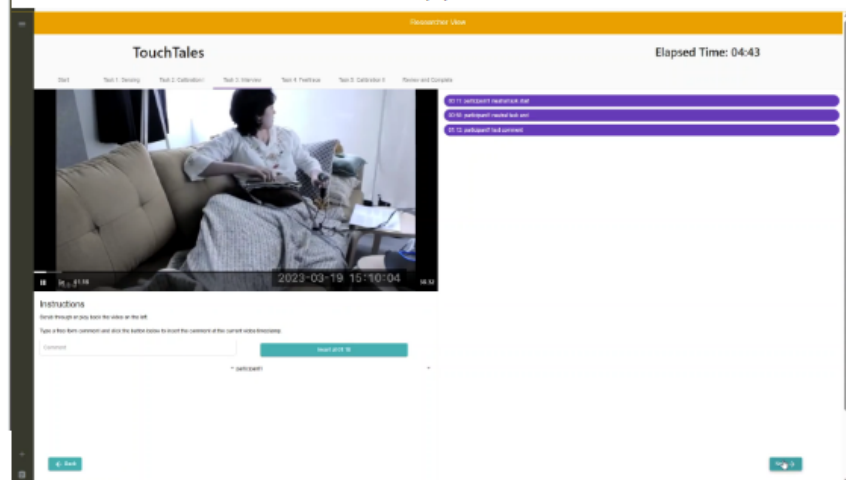
Researchers can easily modify modules including label ranking through drag and drop and addition of multimedia stimuli files. Component documentation facilitates customization of input sensor data types and editing of label arrays.



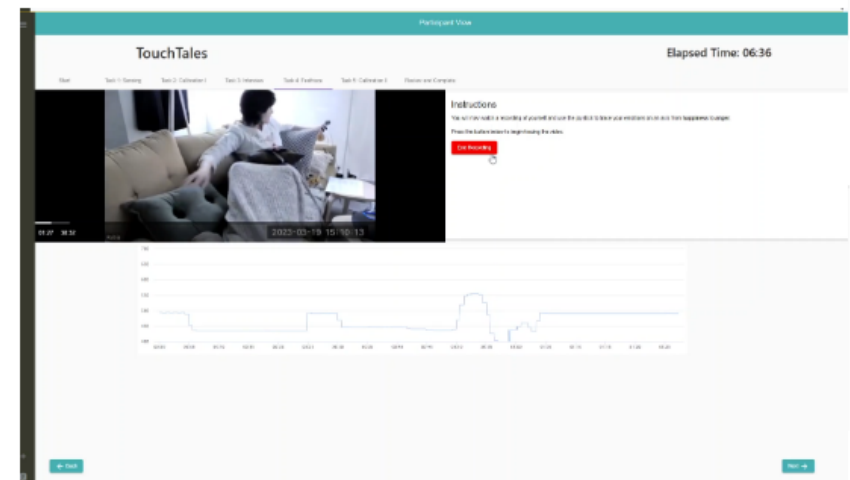
(a)



(b)



(c)



(d)

Figure 8.1: Interface screenshots showing (a) live sensor visualization; (b) emotion word calibration; (c) interview annotation; and (d) continuous annotation stages.

8.3.2 Recruitment Summary

After piloting with two individuals engaged in four sessions, we recorded data from $N = 5$ participants over ten emotion tasks. Due to the sensitive nature of personal stories, we wanted to ensure that (a) participants were comfortable sharing in such a way that it was no worse than venting to or sharing with a friend; and (b) we could reach out to participants to ask how they were feeling a day or so after each session and that they would be able to answer honestly about their ability to manage any trauma, should it be triggered. All participants were known to researchers. Our research team does not include trained psychologists so we selected participants based on who would be comfortable being open about private and sometimes deeply unpleasant moments from their lives.

8.3.3 Ongoing Consent Practice

We posit that differentiating between labelled data instances is easier when emotion expressions are strong and distinct. In order to elicit strong emotions, we ask participants to emote through and describe important and powerful moments from their lives. We learned from our experiences with participant storytelling in [42] (Chapter 5) to prepare for sensitive and emotionally fraught experiences and prioritized participants' sense of comfort by outlining three tactics (determined from our experiences with piloting and discussions with clinicians): (1) an "ask twice" consent process wherein researchers would only ask twice about an emotional element and if the participant was not forthcoming, researchers would change the course of the elicitation intent; (2) participants are encouraged to share and express as they wish including cursing, yelling, singing, over- or under-sharing, whatever allows them to authentically experience their feelings in an honest manner; (3) we reviewed collected data with participants, allowing them to decide the level of privacy they are comfortable at each stage (for example, P1 consented to publication of their emotion label data but their original story was to be analyzed only be those present during data collection; P4 consented to public release of anonymized story text for their second session but not first or third); and (4) after the emotional storytelling task, we gave participants a chance to decompress as appropriate, and scheduled a time within 48 hours for a check in.

8.3.4 Setup

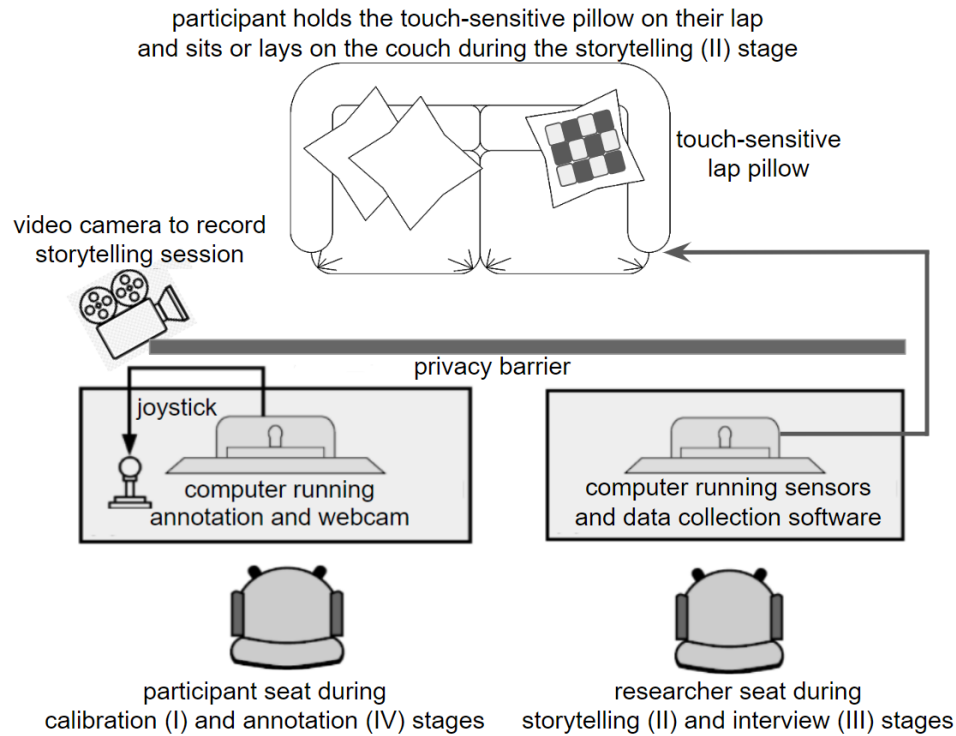


Figure 8.2: Physical setup showing room layout and relative positioning for participant and researcher over all stages of data collection.

For every session, we collected biometric data (heart rate and skin conductivity) and touch pressure; these data streams were synchronized to the storytelling recording (video and audio). In an experiment space furnished to resemble a homey living room, we instructed participants to sit or lay down on a couch, whatever felt more comfortable. Heart rate and GSR sensors were placed on their non-dominant hand; their dominant hand remained on a touch-sensitive pillow that can be positioned either on their lap or chest, depending on their body position. A large opaque separator was positioned in front of the couch to obscure participants' view of the rest of the room. Although participants were alerted to the videocamera at the side of the room, the divider restored some privacy and helped keep focus on the story and resulting feelings without the distraction of the researcher or other

setup details.

8.3.5 Data Collection Protocol

We followed the four-stage multipass emotion data collection structure described in [44, 46]: I. Primary Emotion Task, II. Emotion Word Calibration, III. Interview, and IV. Continuous Annotation, with some customizations. First, we used storytelling as the Primary Emotion Task in order to collect naturalistic expression of emotional touch behaviour (rather than incidental touch pressure emerging from pressing on video game control keys as in [44]). Second, we allowed participants to define their emotion scales based on the most prominent feelings they experienced during the Storytelling stage since we could not know for certain what emotions would be elicited. Lastly, we added a fifth stage: V. Post-Task Reflection to better understand the participant experience in generating this data.

I. Storytelling

As a preliminary step, we asked participants about their present emotions, noting the existence of any holdover emotions from earlier in the day or anything extraordinary that might present a challenge to honest expression. For all 10 sessions, participants reported feeling “fine” or “ready” before hearing the story prompt.

We prepared a set of prompts to help participants hone in on strong emotional stories as shown in Table 8.1. These prompts are counterbalanced to encourage rough coverage of positive- and negative-valenced stories.

Participants were given as much time as was needed to fully explore their prompts. A researcher (who is known to the participant) would ask semi-structured interview questions to clarify context (“What did that mean to you?”), probe for emotional content (“How did that make you feel?”), or engage in conversation to extract deeper emotional memories (“Was that the most difficult/painful/happiest part?”). The storytelling session ended when the participant was not adding more to the story spontaneously and the researcher had no more questions rooted in the original story, or the participant indicated they had completed their story (e.g., “And that’s pretty much it.” – P4-1).

Table 8.1: Storytelling Prompts and the Most Prominent Emotions Elicited as Reported by Participants. Mean(SD) duration of all stories is 9:30 (4:18).

Prompt	Reported Emotions	Opposites	Duration
What are you most proud of?	Pride, Contentment	Shame	4:48
Tell me about the person you share news with first.	Love, Pride, Gratitude	Sadness, Loss	8:53
What is the nicest thing that's been done for you?	Pride, Accomplished	Doubt, Helplessness	5:57
When did you feel the most satisfied with your life?	Connectedness, Longing	Loneliness, Anger	7:12
What do you remember about the best period of your life?	Excitement, Nostalgia	Shame, Embarrassment	7:00
Mean(SD)			6:46 (1:31)
What is your biggest fear?	Anxiety, Dread	Accomplished, Fulfilled	10:55
What is the biggest stressor in your most important relationship?	Guilt, Sadness	Fulfillment, Satisfaction	13:15
What was the hardest decision you've ever made?	Anxiety, Fear, Gratitude	Disgust	7:43
What is your biggest frustration?	Anxiety, Confusion	Satisfaction, Gratitude	9:40
What is your current biggest worry?	Regret, Longing	Anger, Sadness, Spite	19:33
Mean(SD)			12:13 (4:33)

II. Emotion Word Calibration

To understand the emotional landscape of the story content, participants were asked about the most prominent emotion experienced during storytelling, marking the top end of a vertical scale. The participant-perceived opposite of that scale defines the bottom end. For example, if a story's most prominent emotion was "Caring", the opposite end might be labelled "Indifference". If participants had a hard time determining an opposite, researchers suggested words until participants were satisfied with the scale definition.

From a pre-defined list of 12 emotion words – "fear", "love", "sadness", "happiness", "disgust", "surprise", "embarrassment", "envy", "pride", "sympathy", "gratitude", and "anger" – participants were instructed to drag-and-drop as many as they wished to a location on their vertical scale as determined by the prominent emotions from the **I. Storytelling** stage (see Figure 8.1 - top right). The set of words

calibrated along the emotion scale is heretofore referred to as *Calibrated Words*.

III. Interview

To ensure a consistent timeline relative to the original emotion expression, participants reviewed the video recording from the **I. Storytelling** stage with a researcher. The researcher paused the video and annotated the event timestamp with participant comments based on emotionally poignant moments, explanations for their expressions or verbalizations, unusual or unexpected behaviours (e.g., sudden laughter in an otherwise sad story), sudden breaks in prose, etc. During this stage, researchers asked about elements of the story and participants added other details; the interview stage often felt conversational with participant background commentary annotated at storytelling timestamps. By introducing emotion words in **II. Calibration**, we intentionally primed participants to use these words or common synonyms in their descriptions here.

The outcome of the interview stage is a timeseries of phrases and comment annotations aligned with the **I. Storytelling** stage, including touch behaviour, GSR, and HR data. By aligning participant use of Calibrated Words (or synonyms as determined using Python’s Natural Language Toolkit or NLTK library¹) with the story timeline, we generated a timestamped set of *Calibrated Words*. Each data-point (x,y) consists of x = timestamp along the **I. Storytelling** timeline, and y = the vertical distance along the emotion scale of the *Calibrated Word* used during the **III. Interview**. This Timeline with *Calibrated Words* or (Timeline with Calibrated Words (TWCW)) can be used as a supplementary labelset for classifying the modality emotion data collected during the associated timepoint.

IV. Continuous Annotation

As a final labelling stage, participants are moved to a computer station to rewatch the original video recording from the **I. Storytelling** stage through without pausing. Using a custom-built unbiased 1-D joystick [43, 44], participants annotate their emotion experience of the **I. Storytelling** by moving the joystick up/down along the indicated emotion scale as determined in **II. Calibration**. They are instructed

¹Detailed info found at <https://www.nltk.org/>

that moving the joystick up or away from themselves moves the corresponding marker up towards the top of the vertical scale and so moving the joystick down or toward themselves moves the marker down to the low end of the scale. The marker “draws” a continuous line so holding the joystick still shows a horizontal line, suggesting a relatively constant emotion state for the duration of the hold.

The resultant dataset is a continuous timeseries where the x-axis is the timeline from the **I. Storytelling** video and the y-axis is the emotion scale defined in the **II. Calibration** stage.

V. Post-Task Reflection

To close the session, we asked a few procedure reflection questions to assess how participants felt throughout the process, the intensity of their emotion recall, and how they interpreted the emotion scale across labelling stages. We also used this opportunity to ask how stable they felt before leaving the study, checked if they were still comfortable with their earlier data access permissions, and verified that we could check in with them within 48 hours, particularly if they seemed distraught during the storytelling and interview stages.

8.3.6 Dataset Description

The data collected at each stage consists of:

- (I) a video recording and a primary timeline for synchronization, emotion data in the form of GSR, HR, and touch behaviour from interactions with a touch-sensitive pillow;
- (II) a set of *Calibrated Words*;
- (III) rich comments and details describing the emotional experience of recalling and telling a personal story;
- (II+III) a labelset consisting of a timeline of *Calibrated Words* (and synonyms) that describes the primary emotional experience due to the storytelling task that may be the driver of the captured emotion in the HR, GSR, and touch behaviour data – herein referred to as the *Timeline with Calibrated Words* or TWCW.

- (IV) a continuous annotation forming a timeseries of the emotional experience from the storytelling on the emotion scale defined in **II. Calibration**
- (V) an intra-rater reliability check on emotion rating stability

From this dataset, we perform feature extraction on the primary emotion data (namely the GSR, HR, and touch behaviour) and label extraction on the Continuous Annotation data and the TwCW. Combining and windowing these generates labelled data instances that we can use as training data for machine learning classification models.

8.4 Participant Experience

To better understand the data collection and labelling experience, we asked participants for their reflections on the protocol and earmarked interesting behaviour and preferences.

8.4.1 Consent Process

As the story prompts inspire personal tales from participants' lives, we constructed a rather comprehensive consent process in the hopes that it would keep everyone feeling comfortable through multiple sessions. However, from post-session interviews, consultations with expert researchers, and our own team discussions, we noted that the consent process does not need to be a heavy process. In fact, both single and multi-session participants were comfortable with the storytelling procedure. They noted that they liked the consent check-ins over sessions but preferred to keep it short in favour of a shorter protocol.

For participants speaking honestly, knowing what data is being recorded and how it will be used is valuable. In one notable session, our participant decided at the close of the session that they were uncomfortable with their story being reviewed and analyzed, indicating a preference for us to delete the story recordings completely so we deleted the video recording (and backup) in front of them, retaining only innocuous quantitative records (calibrations and continuous annotation) Going forward, our consent process will be shortened to focus on the review of what is being recorded but will still include check-ins both prior to and after each

session. By highlighting which datastreams are being recorded and how it would be used in analysis or presentation allows participants to decide their comfort level with storytelling and future use.

8.4.2 Observations

We note how participants describe the emotions that arise in storytelling and labelling.

Emotions in Storytelling

Emotions generated from storytelling prompts are not always predictable. Participants' lives are personal and varied and prompts elicited many distinct emotions, intertwined in complex ways. For each story, we asked about the most prominent emotion that dominated as the relived emotion in the session. Participants named more than one in every case; sometimes describing concurrent emotions of contrasting valence. When describing the hardest decision they'd made, one participant reported feeling "Anxiety", "Fear", and "Gratitude" when asked for the strongest or most prominent relived emotion.

Overall, the positive-valenced emotions we heard included Pride, Contentment, Love, Gratitude, Accomplishment, Connectedness, Longing, Excitement, and Nostalgia. Negative emotions included Anxiety, Dread, Guilt, Sadness, Fear, Confusion, and Regret. A Student's t-test of the story duration found that negative-valenced stories were longer (significant at $p < 0.05$, large effect size at Cohen's $d = 1.60$) than the positive-valenced stories.

Emotion Scale for Continuous Annotation

Strategies for continuous annotation ranged from moving the joystick "*step-wise*" (P5) to "*crank[ing] up or down for extreme emotions*" but "*reset[ting] to neutralish*" (P4). P1 noted that they were "*playing with the range at first*" before "*reliv[ing] the experience to feel the matching trace*".

When referencing the emotion scale during continuous annotation, there was a range of approaches. P2, P3, and P4 all commented on the scale as intensity, thinking of the scale as "*how intensely did I feel [the emotion at the top]*" (P3). P1 also treated the scale as having "*two ends ... like the intensity of fear vs pride*" but also added that they "*treated the scale more like a binary rather than a spectrum*".

[and] felt like it was a little bit bouncing across a binary.” In contrast, P5 felt that the scale “*could have been more like generically positive or negative feelings*” suggesting that they may have looked at it more like a valence-scale. However, P5 also indicated that they “*head[ed] to the end of the scale for stronger emotions [and] stayed middling otherwise*” which suggests a more intensity-like use that is very similar to P1’s treatment.

8.4.3 Questionnaire Responses

At the end of the storytelling stage, we asked participants to reflect on three questions about the emotions that they felt while reliving their memory. On a 10-point Likert scale, they rated

1. **Relived Emotion Similarity** or how close the relived emotions were to the actual occurrence (1-Not at all; 10-Perfect Match)
2. **Relived Emotion Intensity** or how intense the relived emotions were to the actual occurrence (1-Not at all; 10-Perfect Match)
3. **Relived Emotion Resolution** or how resolved the events or emotions are at this time (1-Currently Active; 10-Made My Peace)

Running Pearson’s correlation on these ratings (see Table 8.2), we find

- **Relived Emotion Similarity** and **Relived Emotion Intensity** to be positively correlated at **0.632**
- **Relived Emotion Similarity** and **Relived Emotion Resolution** had a weak negative correlation at **-0.475**
- **Relived Emotion Intensity** and **Relived Emotion Resolution** were also negatively correlated at **-0.611**

It appears that relived emotions are most similar in kind and intensity to their original occurrence when the events of the story are not yet resolved (for P2, P4-3, P5-1) or where the feelings from those stories are recurrent, even though some of these stories originally occurred in childhood many years ago (for P3-2 and P4-1). In these cases, participants may have been grappling with active feelings during their retelling.

Table 8.2: Relived Emotion Ratings. Participants with multiple sessions are denoted P#-S where S is the session number (P3-1 indicates participant 3’s first storytelling session).

P#	Similarity 1-Not at All 10-Perfect Match	Intensity 1-Not at All 10-Perfect Match	Resolution 1-Currently Active 10-Made My Peace
P4-1	10	9	1
P4-3	10	9	1
P5-3	6	4	1.5
P2	7	8	2
P3-2	6	6.5	2.5
P5-1	3	7	2.5
P3-1	8.5	4	3
P1	4	5	4.5
P4-2	8	6	8
P5-2	2	2	8
Mean(SD)	6.85 (2.69)	5.95 (2.39)	3.95 (3.36)

8.5 Model Training Summary

As classification is still in progress and the focus of another author, this section summarizes our model training procedure, in order to provide important context to the protocol validation – the primary objective here.

8.5.1 Datastream Pre-processing

We constructed a data processing procedure to systematically organize the physiological data collected in **I. Storytelling** stage. A 1D polynomial Savitzky-Golay filter was applied on both GSR and HR data streams to reduce instrumental noise. In our preprocessing of the Continuous Annotation data, we used an Exponential Weighted Moving Average (EWMA) model with a decay parameter of 0.3 to remove hardware ‘jitter’. We found through trial and error with visual inspection that this value provided a satisfactory balance between responsiveness to recent changes and noise reduction.

8.5.2 Feature Extraction and Selection

We partitioned the time series data into equal-sized windows to facilitate our data analysis. Similar to past investigations [44] on touch pressure and its affects on expression of emotions, we obtained the statistical indicators, including mean, variance, maximum, minimum, area under the curve and sum of absolute differences for GSR, HR and touch data. These comprise the features for each window to represent the participants' touch behaviours and bio-signal fluctuations. Moreover, we performed spectral analysis on the touch and bio-signal data streams to compute features in the frequency domain. We also computed the average sum of the touch values per window as our touch feature to account for large area touch behaviour changes.

8.5.3 Label Extraction

Labels serve as the ground truth for model predictions based on input feature data. To capture participants' emotion experience, the TwCW and joystick values marking emotion evolution over the storytelling timeline (from **II+III Calibrated Words+Interview** and **IV. Continuous Annotation** stages respectively) are used to generate three distinct label types for each session: position, angle, and calibrated words.

The position label represents the location or 'emotion state' that a participant was in within the given window and is derived by finding the mean across the continuous annotation within the window. Across all participants and associated data, normalized position annotations create four equally sized bins for values between [0,1].

Representing the direction of emotion evolution, the slope of the curve is calculated within a given window at fixed intervals, normalized, and digitized into four equal-sized bins for slopes, expressed as angles $(-\frac{\pi}{2}, \frac{\pi}{2})$. These comprise the set of angle labels.

This comprehensive label extraction methodology ensures the robustness and accuracy of the ground truth for subsequent model training and evaluation.

8.5.4 Model Exploration Summary

By using strategic portions of the dataset, we can tease out differences in classification accuracy by modality, label type, window size, emotion ranges, feature importance, and generalized vs personalized classifiers. Reporting will be based on 3-fold cross-validation during training and include confusion matrices, performance scores, and best model parameters by estimator.

We evaluate four distinct estimators – specifically Extra Trees, Random Forest, Ada Boost Classifier, and Gradient Boosting – sourced from the Scikit-Learn ensemble library to explore how classification performance varies over variations on training and test data. To optimize model performance, we determined the number of estimators to be either one or two multiples of the number of features. Specifically, if the dataset comprises N features, our preliminary number of estimators was set to N and $2*N$. For the Gradient Boosting Classifier, we explored learning rates of 0.8 and 1.0. The training process involved the construction of a Scikit-Learn Pipeline, integrating Scikit-Learn RFECV and GridSearchCV. The RFECV class from the Scikit-Learn feature selection library executed recursive feature elimination with cross-validation to identify optimal features, while the GridSearchCV class from the Scikit-Learn model selection library systematically explored parameter values for each estimator (see Figure 8.3). Overall, training and testing of emotion direction (as calculated by the angle of inclination from the slope of the continuous annotation curve) from personalized models resulted in classification accuracies of 58% where chance is 25% – similar in performance to our previous work [44].

By Modality

We break down the classification performance by modality to find the minimally viable combination of touch, GSR and HR. Feature selection can further reduce the required computation for a given performance level. Devices designed for realtime responsiveness may have constrained onboard resources and reducing the sensing and/or computational load can significantly improve latency issues and affordability.

By Label Type

We explore classification accuracy by position, angle, and the combination

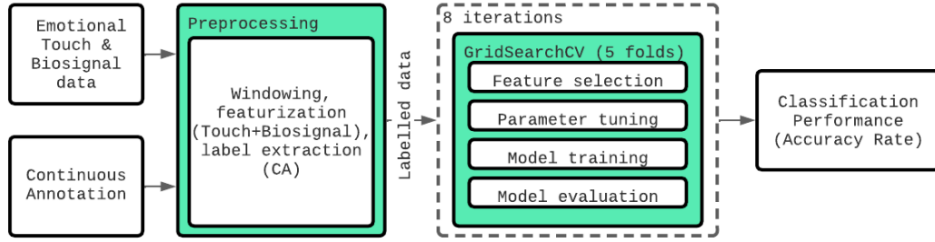


Figure 8.3: Pipeline for model selection and evaluation. We performed grid search CV ($k = 5$) on the training set to tune hyperparameters and select best-fit models for the touch data. The models were then evaluated on an unseen test set to calculate performance metrics. We repeated this process 8 times per participant-session, and report mean test scores across the 8 runs and 10 sessions.

thereof to compare the feasibility of using continuous annotation derived labels for dynamic emotion classification. It may be that high resolution labels are not necessary, in which case we can reduce sampling rate, potentially improving computational efficiency for label generation.

In contrast, we also evaluate the use of calibrated words – a similar class set for emotion state.

Longitudinal Personalized Models

Personalized models have performed well over generalized models in classifying emotion from touch in past work [41, 42]. To simulate realtime classification, we examine how well personalized models trained on past sessions may work on future sessions by experimenting on models trained on *Participant_p*’s first and second sessions and test on their third. We can also compare performance from a generalized model, training on all participants’ sessions and testing on a single session.

Where personalized training continues to yield higher performance than generalized models, we would argue for incorporating individualized training through a similar data collection protocol as described here. Should the performance gain over generalized models be minimal, then it could be argued for devices to be pre-installed with a standard model built on data from many individuals. This approach could improve the immediate out-of-the-box experience, reducing the training ef-

fort by each individual user.

8.6 Limitations and Future Work

We discuss participant reactions and make recommendations on how to improve training procedures for emotionally responsive devices, reflecting on limitations in this proof-of-concept.

8.6.1 More Participant Data

To the best of our knowledge, this is the first training procedure of its kind – building training data for computational models of naturalistic affective touch that evolve with the user. In trying to explore the feasibility of intermittent and repetitious model training of strong emotion expression, we invited a very small number of participants in order to assess and respond to any complications or issues at each stage. We acknowledge that these five users will not be representative of the full scope of experience; nevertheless, we use the lessons of these participants to inform the design of a more comprehensive evaluative study on model development, intended to extend to training interactive devices.

Furthermore, all collection to now has been done in-lab. For a device’s training system to evolve with the user, we know that it must have some reasonable functionality straight out-of-box, and ideally have a relatively seamless in-use training and collection protocol. Consider how your cellular phone’s keyboard has a predictive text system that works well enough globally, but with increased familiarity with your typing and textual composition behaviour, helps you become a much more productive and accurate texter. As another example, you may have had to ‘recalibrate’ your phone’s Map application, periodically having to draw figure eights with your device to re-orient the gyroscope and improve compass accuracy. In either case, some data model is pre-loaded prior to use and works ‘well enough’ such that any one can pick it up and be productive straight away, even if it is not perfectly optimized. Even with a built-in generalized model, our system is likely to require more of a manual alignment as that of the Maps example than the more natural in-use learning seen in autocorrect and predictive text, but we imagine that intermittent requests to ‘recalibrate’ to the user might be welcome to increase the

system’s prediction appropriateness and improve the user experience.

For class annotation, interpretation of the vertical scale during continuous annotation can vary greatly. Joystick movements during Continuous Annotation can represent a range of usages, from that of a binary toggle to an intensity trace. These different strategies may present generalizability concerns. A larger participant pool can also help to identify whether approaches are emotion- or individual-dependent wherein we can better group labelled data to improve generalized models for classification.

8.6.2 Emotion Specificity

To evaluate the protocol and understand the range of emotions that can be elicited in storytelling, we asked very general prompts without expectations of specific emotions. Through 10 prompts, balanced only for negative- and positive-valenced emotions, participants indicated nine distinct dominant emotion words for each valence range (brackets to indicate the number of stories that featured non-unique emotion, 1 otherwise): Pride (3), Contentment, Love, Gratitude, Accomplishment, Connectedness, Longing, Excitement, and Nostalgia for positive prompts and Anxiety (3), Dread, Guilt, Sadness, Fear, Gratitude, Confusion, Regret, and Longing for negative prompts (see Table 8.1).

The approach we chose here focused on the individual experience but it may be valuable to investigate a small set of specific emotions that are easy to evoke and explore how the same emotion manifests for a larger set of individuals. Comparing emotions like Pride and Anxiety – each appearing three times as the dominant positive- and negative-valenced emotion respectively – or Longing and Gratitude – each appearing as both positive and negative prompts – and contrasting the expression evolution of each could help highlight how generalizable emotion expression may be across individuals. For instance, we wonder how Anxiety presents across people and if it might be more consistent compared to something like Longing which may be a little more ambiguous in valence.

8.6.3 Prompting Emotion Evolution

Time is needed to encourage the fullest emotional range through the data collection sessions. From tracking the story durations, we found that the negative-valenced story prompts produced nearly double the time of data collection sessions than the positive-valenced prompts (see Table 8.1).

To better understand the duration discrepancy, we plan to perform semantic analysis on the recorded stories, including evaluating story word counts, pauses, direction changes and peak counts in continuous annotation data. We wonder if the extended length for these negative story prompts may be due to participants providing more backstory and justification for the negative feelings they harboured, and/or due to emotional pauses for those who got choked up and paused speaking in order to let their feelings unfold naturally, or perhaps it takes longer for deep feelings to resolve to a satisfying story close, particularly where events or feelings are most active or unresolved.

From observing these longer data collection sessions, we notice elements of cathartic release from participants. We listened as our participants cried, paused to collect themselves, made connections to past or current events that triggered strong responses and gave them time to let these feelings resolve during the session. We noted how some of these stories involved personal reflections that were reminiscent of a think-aloud version of journalling. Research has shown significant benefits to cognitive and emotional processing via journalling [119, 299], whether written or spoken [195]. Interestingly, it may be that confronting or inviting strong negative feelings can introduce a coactivation such that positive feelings are experienced in the aftermath [6]. For devices intended to help with addressing and/or attending to negative feelings, we wonder if such a protocol of expressing and examining strong feelings may have a two-fold benefit: (1) emulate the natural progression data through negative and positive emotions for training data and (2) provide a guided process to help release existing emotional tension [299]; the latter requiring due consultation with clinical experts.

8.6.4 Data Collection and Training in Use

For an emotion-aware therapy robot or other interactive agent to respond seamlessly to evolving user behaviour, the data labelling procedure needs to allow for continuous, unsupervised data collection. This could look like a background process that captures many diverse contexts, prioritizing multi-session variability to improve real-time interactivity and personalized emotion modelling. We can imagine that a similar protocol involving the development of conversational agents may generate and edit instructional texts to assist users in their daily lives, probing for deeper class justification or description much like the researcher/interviewer in this in-lab version. In Chapter 5, we assessed the impact of classification performance after removing recency-based similarities in user behaviour by dropping data instance neighbours (referred to as gapping). From this, we found that these 'gaps' contributed to performance loss, particularly where data count was dramatically reduced. However, in real-world use, models are often built out-of-session, where training data may have been collected days or months earlier. To better approximate this more challenging classification task, we can train on previous sessions and test on a later one. Future work examines more classification performance on longitudinal participants to better understand the effect of model evolution and data incorporation.

8.7 Conclusion

We presented a data collection protocol for naturalistic affective touch arising from authentic and spontaneous emotion expression. The four-stage procedure involves storytelling as the primary emotion elicitation. Participants responded to a simple prompt that encourages a positive or negative dominant emotion. By using a joystick for continuous annotation of the storytelling experience, participants traced an emotion evolution curve to describe the emotion dynamics of the session wherein the horizontal axis is the timescale of the story as in a timeseries. The vertical scale is participant-defined by the most prominent emotion they felt during storytelling to their perception of the opposite emotion.

Overall, we found:

- negative-valenced prompts generated stories nearly double in duration than

positive-valenced prompts

- continuous annotation of emotion evolution over time can have at least three treatments of the vertical emotion scale: as a true spectrum between opposite emotions, a binary toggle, or an intensity scale for a reference emotion. Knowing which approach a participant used may influence how the label is treated in classification.
- emotional stories that are most active and unresolved, no matter how long ago the original event occurred, produced the most intense and emotionally true experiences in retelling.

Reflecting on these outcomes, our future work involves examining the evolution of specific emotions or a limited range of emotions so as to focus the design of personalized devices that can better respond to user characteristics, preferences, and needs under particular circumstances. For instance, if our emotionally responsive devices are intended for management and regulation through negative emotion experiences, personalized training data can use negative-valenced prompts to be better attuned to particular expressions.

With evidence that this guided multistage labeling protocol produces classification rates within the state-of-the-art recognition range for incidental touch modeling (Chapter 7), we are poised to further this line of study towards the development of real-time emotionally reactive machines. The growing interest in naturalistic paradigms, guided by theoretical models, provides a strong foundation for the continued exploration of emotions with naturalistic touch stimuli.

Chapter 9

Conclusions and Reflections

We explored the early stages of building emotionally responsive devices in two parts: (I) through example devices and the roles they take on for the human user(s) and (II) the development of an onboard emotion classification model prioritizing realtime emotion evolution.

As a result of this work, we establish that machines acting in a variety of emotionally interactive roles can produce and extract emotional touch expression. There is an opportunity to enhance this interactivity by embedding such machines with a personalized emotion recognition engine. By periodically revisiting a guided data collection and labelling protocol, we may be able to update a classification model such that it evolves with the user over time. Closing the (affective touch) interaction loop [321] is a challenging design exercise that may require different approaches by use case and feature priority. In the following, we reflect on four topics of future investment that can inform real use scenarios, namely (1) machine responses to users' emotion expression, (2) protecting trust in technology using emotion data of a private and personal nature, (3) facilitating long-term engagement extending beyond curiosity for novel technologies, and (4) examining the real life environments that may best make use of such devices, and (co-)designing for people most interested adopting machines engaging in emotional touch.

9.1 Accounting for the Neurophysiology of Touch

Touch sensory systems develop from neonatal stages where infants develop a sense of touch beginning at 7 weeks gestation within the womb [16, 197]. The importance of physical contact with caregivers in early infancy has been well-established [26]. From work in the effects of maternal touch on infant pain response [167], C-tactile fibers are thought to be responsible for associating positive and pleasant sensations from experiencing gentle, slow touch. This effect is felt even when touch is non-contextualized and purely mechanical (participants feel soft stroking without knowing who or what is performing the touch) [166].

In future work, we aim to build on the recent developments in our understanding of the neurophysiology of touch to design more conscious touch experiences that link body and experience. Leaning into the precognition of affiliative touch may have significant implications for the design of interactive systems that support emotional well-being and social development.

9.2 Designing for Emotion Reactivity

This thesis takes the approach that using machine recognition or classification of emotion expression from user behaviour allows for having more knowledge of the user's current emotional status which in turn improves the likelihood of triggering an appropriate device or robot response. Another approach we could have taken is to assume that users will adapt their emotional reaction and their narrative interpretation to all robot actions.

Jung and Hinds posit that the cultural and social context for interacting with a robot may be key to understanding robot influence on social and emotional perceptions of an individual [147], alleviating the need for human behaviour/emotion classification. This wider-angle perspective is necessary for field studies investigating how robot presence and behaviour may influence a larger social dynamic within specific environments. For robots intended for use as a personal emotion regulation assistant or other one (human)-to-one (robot) emotional engagement interactions, we argue that being able to recognize the individual 'tells' or 'triggers' may help reduce some of the risk of negative interactions or otherwise unproductive human reactions. While we may never eliminate all risk of erroneous classification and

subsequent robot behaviour, we posit that with the right post-error repair intervention, like apologizing [89] and/or verbal [149] or touch-based reassurance [56], the trust or emotional bond can still be recovered. As imperfect as people are at interpreting the emotions of those around us, humans continue to interpret and respond, learning to better ‘read’ each other as time goes on, often building and strengthening relationships despite repeated missteps [99].

Both approaches have great merit and, when used in conjunction, offer a more holistic understanding of the social and emotional influence robot interactors may have on both independent individuals and social groups. While we have focused on naturalistic emotion recognition here, future work includes examining the impact deployed devices may have on the lives of users and their social circles.

9.3 Designing for Accountability and Trust

With advances in generative Artificial Intelligence (AI), there is a heightened awareness of ethical and safety issues whenever users are providing personal and/or intimate data about their private lives. We expect devices purporting to provide intimate emotional help, whether for regulation or outlet, may be privy to raw or volatile moments. In cases of emotion regulation or therapy-aide applications, users and clinicians alike need to be able to trust that user data is safe and secure in order to engage honestly and purposefully. Therefore, the recording, storing, and use of data should be handled with care and transparency. For instance, we prioritize touch as an embedded emotion modality for its relative unobtrusive collection. However, we are also aware that by virtue of not requiring on-body instrumentation for touch collection, it could be easy for users to lack awareness that their touch data is being recorded and tracked.

Technologies that track private citizens are becoming more prevalent and may be downright unethical when deployed for societal control [162, 186]. The devices we are building involve much of the same user data – physiological and behavioural markers, gaze and facial recordings – and we are cognizant that emotion tracking of private individuals can be fraught with risk. While we intend to develop devices for individual human flourishing, we must also be aware of the possibility of irresponsible implementation of technologies and implicit biases that are

not well-understood – e.g., facial recognition being used as predictors of recidivism [83, 144] or, in our case, keypress behaviour tracked in a workplace to assess employee loyalty.

Imbalances in power and data visibility create significant ethical concerns. For our intended care robot applications, users who are patients with impaired cognitive function may not be fully aware of what access they are allowing when engaging with emotionally reactive devices. Patients’ emotions may be the among the last personal elements not broadcasted to care providers; we argue that this loss of privacy should be treated with care.

9.4 Designing for Engagement

The form factor and physical properties of a device can be developed for behavioural affordances (e.g., designing something with a head can produce an attention-orienting response vs. a simple polygon form factor). In order for a sense of ‘liveliness’ or autonomy, robots may need to have some non-determinism in their behaviours to portray that they are independent agents allowed a ‘theory of mind’ [36]. Design is often easier with a palette [49]. If we can create a set of ‘atomic behaviours’ – independent motion building blocks that can be strung together with smooth transitions to construct more complex robot responses – we can quickly produce a large number of distinct robot behaviours. With a well-designed interface, it is possible that the process could be simple enough to allow for direct user involvement.

Personalization of form and function can be an element of user engagement [177], but increased user control in robot design raises a number of interesting questions. First, “how might user involvement at the design level influence user engagement with the robot or the perception of robot autonomy?” and “what are the components of ‘atomic behaviours’ and how should they be constructed?” Early investigation into both questions can involve researcher-led co-design workshops to understand more about the kinds of forms and functions that users imagine they’d want in personal-use robots. Where researchers provide components that are designed to be mutually compatible such that multiple combinations could be formed, users are better able to build novel yet feasible constructions. Our work in creating the

design patterns for the CuddleBit – a highly customizable robot platform for physical emotion display [35] – demonstrated that even simple motions can be highly evocative of emotion with the same programmed behaviour triggering different interpretations when on different physical forms. By allowing users to devise their own form and behaviours [325], motion patterns may emerge that contribute to determining what could be the basic building blocks of ‘atomic behaviours’.

9.5 Designing for Specific Care Contexts

The original intention of this thesis was to include an investigation into a real-world application and use case for a touch-centric emotionally reactive device for a particular care context. Due to the structural complexity and personal nature of the care environments and in order to respectfully support the needs of workers and families and engage in co-design, it was necessary to invest time building trusting relationships, see the use of resources and physical spaces, and understand how this work was done. We invested a great deal of time over 2018 to early 2020 in two different but equally valuable care environments.

We spent time with clinicians and researchers at Vancouver General Hospital’s Willow Pavilion, an acute care ward for senior patients managing dementia, where research was already underway to involve care robots for therapy. Patients at Willow Pavilion had regular contact with a Paro robot [132] where Occupational Therapist (OT)s and nursing staff created a care protocol of using the robot to gently redirect patient emotional distress or confusion to the Paro as an external focus. We spoke with nurses who stated that when patients were interacting with the Paro placed on a table top with both patient and care provider on the same side of the table, the Paro robot was treated as a third interactor. After some time, staff could see a reduction of agitation in patients who would then allow staff to perform necessary care activities (administer medication, draw blood or take vital sign measurements) that might have been met with resistance otherwise.

We were inspired to define use cases and design opportunities for devices that enhance and enrich human-human interactions, where some of those interactions could also be self-reflexive. Mario, an older adult managing the early stages of dementia, had some experience with the Paro and found it cute but what he really

wanted was a device that could accompany him at home while he could still live at home with the care his wife. He was interested in a smaller, unobtrusive robot that could fit on a nightstand or counter that could provide him comfort and give him more confidence to live independently as long as possible. He was looking for something that offered practical functions like audible reminders of when to take his medication or go to appointments but also more sentimental ones like being a focus for *his* care. Mario wanted to devote energy into caring for something similar to the way that he once did with a beloved pet, but without the disastrous consequences should he forget to feed or clean it some days – he wanted to avoid this ‘indulgence’ (as he put it) to be another source of tasks managed by his wife and family.

Concurrently, we partnered with Canuck Place Children’s Hospice, embedding ourselves as researchers. We interviewed and observed day staff: nurses, recreational therapists, occupational therapists, and teachers who provided comprehensive care for the families. We also met the children with complicated care needs, the surviving siblings who sometimes travelled long distances and missed their own activities to accompany their families, and the guardians and parents who managed to juggle complex and conflicting priorities for important respite care.

I still regularly think of Jonah, a (then) 8-year old boy with minimal motor control. His family and care team were able to understand Jonah based on some very small movements from three fingers in the middle of his left hand – a relatively recent development at the time. His recreational therapist, Lisa, knew that Jonah liked the counter-pressure of firm hugs and relatively heavy touches, informing us to hold his hand tightly when we read to him and to use consistent and firm pressure whenever we were in contact. Light touches were less perceptible and the tentativeness irritating whereas firmer pressure communicated intent, would get his attention, and provided comfort in companionship as it was not easy for him to request touch. Lisa wanted a device for Jonah that could be reminiscent of a blanket or tube snake-like form factor, something that could sit in the wheelchair with him or that he could lay on or under. She would have liked something with an embedded motor or pneumatic pump that would emulate relatively strong breathing-like motions against Jonah’s body to provide a sense of live physical presence, particularly for his overnight care. She was intrigued by the thought of using physiological

and touch sensing that might allow Jonah to initiate emotion communication – up to this point, caregivers would ask mostly yes or no questions about his emotional and general comfort and he would lift fingers to indicate agreement or not. If we were able to create a model of Jonah’s emotions and a device that would report changes, Jonah could use it as an additional communication avenue, potentially increasing the frequency and range of social and emotional interactions.

Overall, our goal is not to replace human interactions or reflections, but to lower the barrier for social and emotional engagement. Much like how the use of a Fitbit activity tracker might offer reminders and improve motivation to go for regular runs but does nothing to reduce the necessary human work of running, we imagine automatic emotion trace might lower the barrier for tracking or collecting data but would still require the emotional regulation and reflection work to be effective. This thesis begins the work of investigating how to track time-varying emotion trajectories to offer insight into and visualization opportunities for how our emotions evolve spontaneously and authentically under naturalistic conditions, and presents a proof-of-concept for generating evolving personalizable models that could be embedded into emotionally reactive devices. Models like these depend crucially on data collection and labelling, which can be challenging for anyone. To serve Mario and Jonah, who may not be able to consistently provide ground truth labels themselves, we can tap into the experiences of close caregivers like parents and care staff. People who recognize their behaviours or triggers and can act as proxies to report and associate labels on their behalf – a method shown to have similar performance to self-reporting [324].

For the safety of vulnerable patients and care staff, this line of inquiry was paused during the COVID-19 pandemic.

9.6 Summary

In Chapter 1, we posed the overarching question: “*How can we enable machines to recognize true and spontaneous evolving emotion expressed through touch?*”, and produced a set of contributions. Over the course of this thesis we have examined the roles that machines play in emotional interaction with human users and explored emotion recognition engines that evolve with spontaneous user expression. We

close by revisiting the contributions that this journey has yielded.

1. While machine recognition of emotion intent in human-designed haptic messages exceed chance, human users are better at interpreting the same messages but only when they have valuable context clues and relationship history. Augmenting human users with machine classification or ‘prediction’ of transmitted messages may be able to narrow down the possibilities and improve the efficacy of the device as a conduit for machine-mediated emotion intent between users (Chapter 3).
2. Holding pet-sized robots exhibiting ‘breathing’ behaviours can be shown to calm or agitate by varying the frequency and regularity of the waveform (Chapter 4).
3. Telling emotional stories from one’s personal life can produce machine recognizable affective touch expression (Chapter 5).
4. Emotion expression should have a variety of representations, including one that acknowledges the time-varying experience of emotions (Chapter 2).
5. One way to track emotion as it evolves is to review and appraise an experience more than once. While multiple labelling passes create more opportunities for label mismatch, it also increases the likelihood of capturing complex nuance in fast-evolving emotion (Chapter 6).
6. Through a novel dataset of brain activity data and incidental touch pressure collected while participants played a horror video game (the FEEL dataset), we benchmark classification performance of dynamic emotion – evolving emotion within a time window (e.g., differentiating between happy-getting-happier *vs.* happy-getting-anxious) – to F1-scores of up to 0.82 from incidental emotion via keypress force data (Chapter 7).
7. Guided emotion labelling with a custom tool can generate natural and spontaneous emotion evolution training data required for classification models of real-world ‘in-the-wild’ emotion evolution (Chapter 8).

Armed with new techniques for dynamic emotion modelling, we are now in a better position to build the custom, co-designed emotionally reactive devices of our imaginations.

Bibliography

- [1] M. Z. I. Ahmed, N. Sinha, S. Phadikar, and E. Ghaderpour. Automated Feature Extraction on AsMap for Emotion Classification Using EEG. *Sensors*, 22(6):2346, Mar. 2022. ISSN 1424-8220. doi:10.3390/s22062346. URL <https://www.mdpi.com/1424-8220/22/6/2346>. → pages xxvi, 159, 167, 168
- [2] K. Alarabi Aljribi. A comparative analysis of frequency bands in eeg based emotion recognition system. In *The 7th Int'l Conf on Engineering & MIS 2021*, 2021. → page 162
- [3] S. M. Alarcao and M. J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Trans on Affective Computing*, 10(3):374–393, 2017. → pages 13, 156
- [4] J. Allen, L. Cang, M. Phan-Ba, A. Strang, and K. MacLean. Introducing the cuddlebot: A robot that responds to touch gestures. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts*, pages 295–295, 2015. → page 5
- [5] K. Altun and K. E. MacLean. Recognizing affect in human touch of a robot. *Pattern Recognition Letters*, 66:31–40, 2015. → pages 9, 11, 12, 98, 99, 103, 104, 107, 110, 111, 112, 116, 117, 126, 127
- [6] E. B. Andrade and J. B. Cohen. On the consumption of negative feelings. *Journal of Consumer Research*, 34(3):283–300, 2007. → page 199
- [7] B. App et al. Nonverbal channel use in communication of emotion: how may depend on why. *Emotion*, 11(3):603, 2011. → page 156
- [8] B. M. Appelhans and L. J. Luecken. Heart rate variability as an index of regulated emotional responding. *Review of general psychology*, 10(3): 229–240, 2006. → page 181

- [9] B. M. Appelhans and L. J. Luecken. Heart rate variability and pain: associations of two interrelated homeostatic processes. *Biological Psychology*, 77(2):174–182, 2008. → page 105
- [10] A. Asadi, O. Niebuhr, J. Jørgensen, and K. Fischer. Inducing changes in breathing patterns using a soft robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 683–687. IEEE, 2022. → pages 80, 93, 94
- [11] J. N. Bailenson, N. Yee, S. Brave, D. Merget, and D. Koslow. Virtual interpersonal touch: expressing and recognizing emotions through haptic devices. *Human–Computer Interaction*, 22(3):325–353, 2007. → pages 1, 46, 71
- [12] K. Bakhtiyari and H. Husain. Fuzzy model of dominance emotions in affective computing. *Neural Computing and Applications*, 25(6): 1467–1477, 10 2014. ISSN 0941-0643. doi:10.1007/s00521-014-1637-6. → pages 28, 31
- [13] I. Bakker, T. Van der Voordt, P. Vink, and J. De Boon. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3): 405–421, 2014. → page 139
- [14] M. R. Banks and W. A. Banks. The effects of animal-assisted therapy on loneliness in an elderly population in long-term care facilities. *Journals of Gerontology: Biological & Medical Sciences*, 57(7):M428–M432, 2002. → page 98
- [15] S. B. Barker and K. S. Dawson. The effects of animal-assisted therapy on anxiety ratings of hospitalized psychiatric patients. *Psychiatric Services*, 1998. → page 98
- [16] L. Barnett. Keep in touch: The importance of touch in infant development. *Infant Observation*, 8(2):115–123, 2005. → pages 1, 4, 41, 203
- [17] L. F. Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017. → pages 4, 15
- [18] L. F. Barrett and E. A. Kensinger. Context is routinely encoded during emotion perception. *Psychological science*, 21(4):595–599, 2010. → pages 12, 42, 46
- [19] L. F. Barrett and J. A. Russell. *The psychological construction of emotion*. Guilford Publications, 2014. → pages 26, 34

- [20] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross. The experience of emotion. *Annu. Rev. Psychol.*, 58:373–403, 2007. → pages 15, 136, 159, 179
- [21] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011. → pages 42, 69
- [22] R. F. Baumeister and M. R. Leary. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Interpersonal development*, pages 57–89, 2017. → page 1
- [23] E. E. Bernstein, J. E. Curtiss, G. W. Wu, P. J. Barreira, and R. J. McNally. Exercise and emotion dynamics: An experience sampling study. *Emotion*, 19(4):637, 2019. → page 141
- [24] E. Bhuwalka, Kunal; Icel, Nur; Gong. How does Robot Feedback Affect Participant Affinity and Trust ? *HRI*, 2018. → page 31
- [25] T. Bi, R. Andrea Buono, T. Olugbade, A. Singh, C. Holloway, E. Costanza, A. C de C Williams, N. E. Gold, and N. Berthouze. Towards chatbot-supported self-reporting for increased reliability and richness of ground truth for automatic pain recognition: Reflections on long-distance runners and people with chronic pain. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 43–53, 2021. → page 16
- [26] A. E. Bigelow and L. R. Williams. To have and to hold: Effects of physical contact on infants and their caregivers, 2020. → page 203
- [27] K. Boehner, R. Roge ´rio Depaula, P. Dourish, and P. Sengers. How emotion is made and measured. *Int. J. Human-Computer Studies*, 65: 275–291, 2007. doi:10.1016/j.ijhcs.2006.11.016. URL www.elsevier.com/locate/ijhcs. → page 33
- [28] F. A. Boiten, N. H. Frijda, and C. J. Wientjes. Emotions and respiratory patterns: review and critical analysis. *International journal of psychophysiology*, 17(2):103–128, 1994. → pages 80, 85, 94
- [29] S. Bonaccio, J. O’Reilly, S. L. O’Sullivan, and F. Chiocchio. Nonverbal behavior and communication in the workplace: A review and an agenda for research. *Journal of Management*, 42(5):1044–1074, 2016. → page 1

- [30] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994. → pages 15, 16, 26, 36, 136, 139, 159, 179
- [31] C. Breazeal, N. DePalma, J. Orkin, S. Chernova, and M. Jung. Crowdsourcing Human-Robot Interaction: New Methods and System Evaluation in a Public Environment. *Journal of Human-Robot Interaction*, 2(1):82–111, 2013. doi:10.5898/jhri.2.1.breazeal. → page 31
- [32] H. Breivik, P. Borchgrevink, S. Allen, L. Rosseland, L. Romundstad, E. Breivik Hals, G. Kvarstein, and A. Stubhaug. Assessment of pain. *BJA: British Journal of Anaesthesia*, 101(1):17–24, 2008. → page 32
- [33] M. Bretan, G. Hoffman, and G. Weinberg. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human Computer Studies*, 78:1–16, 2015. ISSN 10959300. doi:10.1016/j.ijhcs.2015.01.006. URL <http://dx.doi.org/10.1016/j.ijhcs.2015.01.006>. → pages 31, 32
- [34] P. Bucci, X. L. Cang, M. Chun, D. Marino, O. Schneider, H. Seifi, and K. MacLean. Cuddlebits: an iterative prototyping platform for complex haptic display. *Eurohaptics Demonstration*, 2016. → pages 83, 86
- [35] P. Bucci, X. L. Cang, A. Valair, D. Marino, L. Tseng, M. Jung, J. Rantala, O. S. Schneider, and K. 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. → pages 31, 32, 80, 83, 108, 206
- [36] P. Bucci, L. X. Cang, L. Zhang, and K. E. MacLean. Is it happy? behavioural and narrative frame complexity impact perceptions of a simple furry robot’s emotions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, Montreal, CAN, 2018. → pages 45, 69, 205
- [37] P. Bucci, L. Zhang, X. L. Cang, and K. 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. → pages xxiii, 28, 31, 32, 107, 108
- [38] P. Bucci et al. Real emotions don’t stand still: Toward ecologically viable representation of affective interaction. In *IEEE Int’l Conf on Affective*

Computing & Intelligent Interaction (ACII), pages 1–7, 2019. → pages 15, 97, 132, 136, 138, 139, 159, 163, 172

- [39] P. H. Bucci, X. L. Cang, H. Mah, L. Rodgers, and K. E. MacLean. Real emotions don't stand still: Toward ecologically viable representation of affective interaction. In *2019 8th Intl Conf on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019. → pages vii, 16, 42, 44, 180
- [40] R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas. *The Oxford Handbook of Affective Computing*. Oxford University Press, 2014. → page 129
- [41] X. L. Cang, P. Bucci, A. Strang, J. Allen, K. MacLean, and H. S. Liu. Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In *Proc of the 2015 ACM on Intl Conf on Multimodal Interaction*, pages 147–154, 2015. → pages 13, 14, 65, 97, 99, 100, 106, 110, 112, 116, 118, 139, 180, 196
- [42] X. L. Cang, P. Bucci, J. Rantala, and K. Maclean. Discerning affect from touch and gaze during interaction with a robot pet. *IEEE Transactions on Affective Computing*, 2021. → pages vii, 13, 44, 59, 65, 73, 136, 141, 142, 158, 159, 164, 166, 167, 171, 177, 180, 184, 196
- [43] X. L. Cang, R. R. Guerra, P. Bucci, B. Guta, K. MacLean, L. Rodgers, H. Mah, S. Hsu, Q. Feng, C. Zhang, et al. Choose or fuse: Enriching data views with multi-label emotion dynamics. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2022. → pages 177, 180, 182, 188
- [44] X. L. Cang, R. R. Guerra, B. Guta, P. Bucci, L. Rodgers, H. Mah, Q. Feng, A. Agrawal, and K. E. MacLean. Feeling (key) pressed: Implicit touch pressure bests brain activity in modelling emotion dynamics in the space between stressed and relaxed. *IEEE Transactions on Haptics*, pages 1–8, 2023. [doi:10.1109/TOH.2023.3308059](https://doi.org/10.1109/TOH.2023.3308059). → pages 14, 180, 181, 186, 188, 194, 195
- [45] X. L. Cang, A. Israr, and K. E. MacLean. When is a haptic message like an inside joke? digitally mediated emotive communication builds on shared history. *IEEE Transactions on Affective Computing*, 14(1):732–746, 2023. → pages vi, 20
- [46] X. L. Cang et al. Choose or fuse: Enriching data views with multi-label emotion dynamics. In *IEEE 10th Int'l Conf on Affective Computing &*

Intelligent Interaction (ACII), 2022. → pages
156, 157, 160, 161, 163, 165, 173, 181, 186

- [47] T. Carter, S. A. Seah, B. Long, B. Drinkwater, and S. Subramanian. Ultrahaptics: multi-point mid-air haptic feedback for touch surfaces. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 505–514, 2013. → page 11
- [48] D. Caruelle, A. Gustafsson, P. Shams, and L. Lervik-Olsen. The use of electrodermal activity (eda) measurement to understand consumer emotions—a literature review and a call for action. *Journal of Business Research*, 104:146–160, 2019. → page 13
- [49] H. Chang, O. Fried, Y. Liu, S. DiVerdi, and A. Finkelstein. Palette-based photo recoloring. *ACM Trans. Graph.*, 34(4):139–1, 2015. → page 205
- [50] H.-Y. Chen, J. Santos, M. Graves, K. Kim, and H. Z. Tan. Tactor localization at the wrist. In *Intl Conf on Human Haptic Sensing and Touch Enabled Computer Applications*, pages 209–218. Springer, 2008. → page 49
- [51] J. Chen, P. Zhang, Z. Mao, Y. Huang, D. Jiang, and Y. Zhang. Accurate eeg-based emotion recognition on combined features using deep convolutional neural networks. *IEEE Access*, 7:44317–44328, 2019. → page 168
- [52] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proc of the 22nd acm sigkdd Int’l Conf on knowledge discovery and data mining*, pages 785–794, 2016. → pages 164, 167
- [53] H. Choi, D. Brouwer, M. A. Lin, K. T. Yoshida, C. Rognon, B. Stephens-Fripp, A. M. Okamura, and M. R. Cutkosky. Deep learning classification of touch gestures using distributed normal and shear force. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3659–3665. IEEE, 2022. → page 13
- [54] R. W. Cholewiak and A. A. Collins. Vibrotactile localization on the arm: Effects of place, space, and age. *Perception & psychophysics*, 65(7): 1058–1077, 2003. → page 45
- [55] C. Classen. *The deepest sense: A cultural history of touch*. University of Illinois Press, 2012. → page 46

- [56] H. Claire, N. Khojasteh, H. Tennent, and M. Jung. Using expectancy violations theory to understand robot touch interpretation. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 163–165, 2020. → pages 179, 204
- [57] J. A. Coan and J. J. Allen. *Handbook of emotion elicitation and assessment*. Oxford university press, 2007. → pages 9, 12, 102, 110, 129, 142, 180
- [58] J. A. Coan, H. S. Schaefer, and R. J. Davidson. Lending a hand: Social regulation of the neural response to threat. *Psychological Science*, 17(12): 1032–1039, 2006. → pages 78, 79
- [59] C. Conati. Probabilistic assessment of user’s emotions in educational games. *Applied Artificial Intelligence*, 16(7-8):555–575, 2002. → page 141
- [60] C. Conati, R. Chabbal, and H. Maclaren. A study on using biometric sensors for monitoring user emotions in educational games. In *Workshop on assessing and adapting to user attitudes and affect: Why, when and how*, 2003. → page 14
- [61] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder. ‘feeltrace’: An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000. → pages 139, 144, 176
- [62] A. Cranny-Francis. Semefulness: a social semiotics of touch. *Social Semiotics*, 21(4):463–481, 2011. → page 20
- [63] J. R. Crawford and J. D. Henry. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43(3): 245–265, 2004. → page 103
- [64] O. Damm, K. Dreier, F. Hegel, P. Jaecks, P. Stenneken, B. Wrede, and M. 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. → page 31
- [65] B. De Gelder, J. Snyder, D. Greve, G. Gerard, and N. Hadjikhani. Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body. *Proceedings of the National Academy of Sciences*, 101(47):16701–16706, 2004. → page 80

- [66] R. T. Dean and W. Dunsmuir. Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior Research Methods*, 48(2):783–802, 2016. → page 148
- [67] A. Debrot, D. Schoebi, M. Perrez, and A. B. Horn. Touch as an interpersonal emotion regulation process in couples’ daily lives: The mediating role of psychological intimacy. *Personality and Social Psychology Bulletin*, 39(10):1373–1385, 2013. → page 79
- [68] A. DeFalco and L. Dolezal. What is affective technotouch (and why does it matter)? *The Senses and Society*, pages 1–7, 2023. → pages 6, 11
- [69] A. Delorme. Eeg is better left alone. *Scientific reports*, 13(1):2372, 2023. → page 162
- [70] C. Deveney and D. Pizzagalli. The cognitive consequences of emotion regulation: an erp investigation. *Psychophysiology*, 45(3), 2008. → pages 159, 164
- [71] M. Dewitte, C. Otten, and L. Walker. Making love in the time of corona—considering relationships in lockdown. *Nature Reviews Urology*, 17(10):547–553, 2020. → pages 6, 41
- [72] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516, 2013. → pages 3, 12
- [73] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th international conference on multimodal interaction*, pages 461–466, 2014. → page 176
- [74] L. L. Di Stasi, A. Catena, J. J. Canas, S. L. Macknik, and S. Martinez-Conde. Saccadic velocity as an arousal index in naturalistic tasks. *Neuroscience & Biobehavioral Reviews*, 37(5):968–975, 2013. → page 105
- [75] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th Int’l IEEE/EMBS Conf on Neural Engineering (NER)*, pages 81–84. IEEE, 2013. → page 167

- [76] B. Dudzik and J. Broekens. A valid self-report is never late, nor is it early: On considering the “right” temporal distance for assessing emotional experience. In *2nd Momentary Emotion Elicitation & Capture Workshop at CHI*, 2021. → pages 141, 160
- [77] M. Egger, M. Ley, and S. Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019. → pages 13, 177
- [78] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. → page 31
- [79] P. Ekman and H. Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979. → page 14
- [80] P. Ekman, R. W. Levenson, and W. V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210, 1983. → pages 12, 99, 104, 142
- [81] C. Epp, M. Lippold, and R. L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proc of the sigchi Conf on human factors in computing systems*, pages 715–724, 2011. → page 13
- [82] J. A. Fadaei, K. Jeanmonod, O. A. Kannape, J. Potheegadoo, H. Bleuler, M. Hara, and O. Blanke. Cogno-vest: a torso-worn, force display to experimentally induce specific hallucinations and related bodily sensations. *IEEE Transactions on Cognitive and Developmental Systems*, 2021. → page 45
- [83] M. M. Farayola, I. Tal, R. Connolly, T. Saber, and M. Bendecheache. Ethics and trustworthiness of ai for predicting the risk of recidivism: A systematic literature review. *Information*, 14(8):426, 2023. → page 205
- [84] J. W. Fernando, Y. Kashima, and S. M. Laham. Alternatives to the fixed-set model: A review of appraisal models of emotion. *Cognition and Emotion*, 31(1):19–32, 2017. → page 15
- [85] L. Findlater and J. Wobbrock. Personalized input: improving ten-finger touchscreen typing through automatic adaptation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 815–824, 2012. → page 177

- [86] K. Fischer, L. C. Jensen, M. Vanessa, and D. Wieschen. Emotion Expression in HRI – When and Why. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 29–38, 2019. → pages 29, 31
- [87] A. Flagg and K. MacLean. Affective touch gesture recognition for a furry zoomorphic machine. In *Proc of the 7th Intl Conf on Tangible, Embedded and Embodied Interaction*, pages 25–32, 2013. → pages 13, 59, 65, 97, 99, 100, 104, 106, 107, 110, 112, 114, 115, 116, 118, 177
- [88] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3):143–166, 2003. → pages 4, 83, 97, 179
- [89] P. Fratzak, Y. M. Goh, P. Kinnell, L. Justham, and A. Soltoggio. Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *International Journal of Industrial Ergonomics*, 82:103078, 2021. → page 204
- [90] J. Frey, M. Grabli, R. Slyper, and J. R. Cauchard. Breeze: Sharing biofeedback through wearable technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018. → page 11
- [91] D. K. Fromme, W. E. Jaynes, D. K. Taylor, E. G. Hanold, J. Daniell, J. R. Rountree, and M. L. Fromme. Nonverbal behavior and attitudes toward touch. *Journal of Nonverbal Behavior*, 13:3–14, 1989. → page 83
- [92] W. A. Fuller. *Intro to statistical time series*. John Wiley & Sons, 2009. → page 148
- [93] Y. Gaffary, J.-C. Martin, and M. Ammi. Haptic expression and perception of spontaneous stress. *IEEE Transactions on Affective Computing*, 11(1): 138–150, 2018. → pages 3, 7, 11, 12, 97, 99, 141
- [94] G. Gao, M. F. Jung, G. Culbertson, S. R. Fussell, M. F. Jung, S. Young Hwang, G. Culbertson, S. R. Fussell, and M. 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. ISSN 2573-0142. doi:10.1145/3134683. URL [{%}0Ahttps://www.researchgate.net/publication/321193673">https://doi.org/10.1145/3134683{ }0Ahttps://www.researchgate.net/publication/321193673](https://doi.org/10.1145/3134683). → page 33

- [95] Y. Gao, N. Bianchi-Berthouze, and H. Meng. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(4):1–30, 2012. → page 159
- [96] Y. F. A. Gaus, T. Olugbade, A. Jan, R. Qin, J. Liu, F. Zhang, H. Meng, and N. Bianchi-Berthouze. Social touch gesture recognition using random forest and boosting on distinct feature sets. In *Proc of the 2015 ACM on Intl Conf on Multimodal Interaction*, pages 399–406, 2015. → pages 65, 99
- [97] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251, 2014. → page 14
- [98] S. Ghosh, S. Sahu, N. Ganguly, B. Mitra, and P. De. Emokey: An emotion-aware smartphone keyboard for mental health monitoring. In *2019 11th Int’l Conf on Communication Systems & Networks (COMSNETS)*, pages 496–499. IEEE, 2019. → page 9
- [99] J. M. Gottman, J. Driver, and A. Tabares. Repair during marital conflict in newlyweds: How couples move from attack–defend to collaboration. *Journal of Family Psychotherapy*, 26(2):85–108, 2015. → page 204
- [100] J. J. Gross. Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Personality & Social Psychology*, 74(1):224, 1998. → pages 136, 141, 159
- [101] J. J. Gross. The emerging field of emotion regulation: an integrative review. *Review of General Psychology*, 2(3):271, 1998. → page 129
- [102] J. J. Gross. *Handbook of emotion regulation*. Guilford publications, 2013. → pages 141, 181
- [103] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995. → pages 84, 141, 142
- [104] Y. Gu, S.-L. Tan, K.-J. Wong, M.-H. R. Ho, and L. Qu. A biometric signature based system for improved emotion recognition using physiological responses from multiple subjects. In *2010 8th IEEE International Conference on Industrial Informatics*, pages 61–66. IEEE, 2010. → pages 14, 177
- [105] L. K. Guerrero and K. Floyd. *Nonverbal communication in close relationships*. Routledge, 2006. → page 1

- [106] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1): 389–422, 2002. → page 168
- [107] A. Haans and W. IJsselsteijn. Mediated social touch: a review of current research and future directions. *Virtual Reality*, 9(2-3):149–159, 2006. → page 44
- [108] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. → page 116
- [109] M. Hamada, B. Zaidan, and A. Zaidan. A systematic review for human eeg brain signals based emotion classification, feature extraction, brain condition, group comparison. *Journal of medical systems*, 42:1–25, 2018. → page 13
- [110] J. T. Hancock, K. Gee, K. Ciaccio, and J. M.-H. Lin. I’m sad you’re sad: emotional contagion in cmc. In *ACM Conf on Computer Supported Cooperative Work (CSCW)*, pages 295–298, 2008. → page 103
- [111] E. Hatfield, L. Bensman, P. D. Thornton, and R. L. Rapson. New perspectives on emotional contagion: A review of classic and recent research on facial mimicry and contagion. *Interpersona: An International Journal on Personal Relationships*, 8(2), 2014. → page 79
- [112] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski. Under pressure: sensing stress of computer users. In *Proc of the SIGCHI Conf on Human factors in computing systems*, pages 51–60, 2014. → pages 156, 158, 159, 171
- [113] M. J. Hertenstein. Touch: Its communicative functions in infancy. *Human Development*, 45(2):70–94, 2002. → pages 4, 13
- [114] M. J. Hertenstein, D. Keltner, B. App, B. A. Bulleit, and A. R. Jaskolka. Touch communicates distinct emotions. *Emotion*, 6(3):528, 2006. → pages 13, 42, 46, 64, 78, 79, 103, 104, 107, 117, 139, 156, 158, 176, 177
- [115] M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner. The communication of emotion via touch. *Emotion*, 9(4):566, 2009. → pages 1, 13, 46, 78, 79, 117

- [116] M. J. Hertenstein et al. The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research. *Genetic, social, & general psychology monographs*, 132(1), 2006. → page 158
- [117] E. H. Hess and S. B. Petrovich. Pupillary behavior in communication. In A. W. Siegman and S. Feldstein, editors, *Nonverbal Behavior and Communication*, pages 327–348. Erlbaum, Hillsdale, NJ, 1972. → page 105
- [118] U. Hess and S. Hareli. The influence of context on emotion recognition in humans. In *2015 11th IEEE Intl Conf and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 3, pages 1–6. IEEE, 2015. → page 46
- [119] R. Hiemstra et al. Uses and benefits of journal writing. *New directions for adult and continuing education*, 2001(90):19, 2001. → page 199
- [120] R. Hill. Perceptual attention in virtual humans: Toward realistic and believable gaze behaviors. In *AAAI Fall Symposium on Simulating Human Agents*, pages 46–52, 2000. → page 105
- [121] W. E. Hipson and S. M. Mohammad. Emotion dynamics in movie dialogues. *Plos one*, 16(9):e0256153, 2021. → page 141
- [122] M. Hladky, T. Schneeberger, and P. Gebhard. Understanding shame signals: Functions of smile and laughter in the context of shame. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–07. IEEE, 2021. → page 15
- [123] J. Hodges. Making it up and making do: Simulation, imagination, and empathic accuracy. *The Handbook of Imagination and Mental Simulation*, pages 281–294, 2008. → page 31
- [124] K. Hoemann, Z. Khan, M. J. Feldman, C. Nielson, M. Devlin, J. Dy, L. F. Barrett, J. B. Wormwood, and K. S. Quigley. Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific reports*, 10(1):1–16, 2020. → page 153
- [125] T. Hollenstein. State space grids. In *State Space Grids*, pages 11–33. Springer, 2013. → page 28

- [126] M. Hoque and R. W. Picard. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 354–359. IEEE, 2011. → page 12
- [127] M. Hoque and R. W. Picard. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *IEEE Face & Gesture*, pages 354–359, 2011. → pages 104, 141
- [128] D. F. Horwitz. *Blackwell’s five-minute veterinary consult clinical companion: canine and feline behavior*. John Wiley & Sons, 2018. → page 80
- [129] D. F. Horwitz and I. Rodan. Behavioral awareness in the feline consultation: Understanding physical and emotional health. *Journal of feline medicine and surgery*, 20(5):423–436, 2018. → page 80
- [130] M. Houben, W. Van Den Noortgate, and P. Kuppens. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological bulletin*, 141(4):901, 2015. → pages 16, 138, 141, 149, 150
- [131] G. Huisman. Social touch technology: a survey of haptic technology for social touch. *IEEE Transactions on Haptics*, 10(3):391–408, 2017. → pages 1, 42, 44
- [132] L. Hung, C. Liu, E. Woldum, A. Au-Yeung, A. Berndt, C. Wallsworth, N. Horne, M. Gregorio, J. Mann, and H. Chaudhury. The benefits of and barriers to using a social robot paro in care settings: a scoping review. *BMC geriatrics*, 19:1–10, 2019. → pages 5, 10, 206
- [133] S. J. Hunt, L. A. Hart, and R. Gomulkiewicz. Role of small animals in social interactions between strangers. *The Journal of Social Psychology*, 132(2):245–256, 1992. → page 83
- [134] J. S. Hunter. The exponentially weighted moving average. *J of quality technology*, 18(4):203–210, 1986. → page 166
- [135] *HRI ’18: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 2018. International Conference on Human-Robot Interaction, ACM. ISBN 978-1-4503-4953-6. → page 31

- [136] A. Israr, S. Zhao, K. Schwalje, R. Klatzky, and J. Lehman. Feel effects: enriching storytelling with haptic feedback. *ACM Transactions on Applied Perception (TAP)*, 11(3):1–17, 2014. → pages 42, 45, 49
- [137] A. Israr, S. Zhao, Z. Schwemler, and A. Fritz. Stereohaptics toolkit for dynamic tactile experiences. In *International Conference on Human-Computer Interaction*, pages 217–232. Springer, 2019. → page 45
- [138] M. Iwasaki and Y. Noguchi. Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements. *Scientific reports*, 6(1):22049, 2016. → page 14
- [139] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *Intelligent Tutoring Systems*, pages 29–38. Springer, 2014. → pages 13, 98, 100, 105, 176
- [140] K. Jensen. Envelope model of isolated musical sounds. In *Proc of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, volume 12. Citeseer, 1999. → pages 167, 171
- [141] S. Jeong, C. Breazeal, D. Logan, and P. Weinstock. Huggable: the impact of embodiment on promoting socio-emotional interactions for young pediatric inpatients. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018. → pages 5, 10
- [142] C. Jewitt, S. Price, J. Steimle, G. Huisman, L. Golmohammadi, N. Pourjafarian, W. Frier, T. Howard, S. Ipakchian, M. Ornati, S. Paneels, and J. Weda. Manifesto for digital social touch in crisis. *Frontiers Special Issue on Social Touch*, 2021. URL <https://doi.org/10.3389/fcomp.2021.754050>. → page 42
- [143] B. Jin and J. F. Pena. Mobile communication in romantic relationships: Mobile phone use, relational uncertainty, love, commitment, and attachment styles. *Communication Reports*, 23(1):39–51, 2010. → page 42
- [144] J.-M. John-Mathews, D. Cardon, and C. Balagué. From reality to world. a critical perspective on ai fairness. *Journal of Business Ethics*, 178(4): 945–959, 2022. → page 205
- [145] C. M. Jones and T. Troen. Biometric valence and arousal recognition. In *Australasian Conf on Computer-Human Interaction: Entertaining User Interfaces*, pages 191–194, 2007. → page 105

- [146] Y. Ju, D. Zheng, D. Hynds, G. Chernyshov, K. Kunze, and K. Minamizawa. Haptic empathy: Conveying emotional meaning through vibrotactile feedback. In *Extended Abstracts of the 2021 CHI Conf on Human Factors in Computing Systems*, pages 1–7, 2021. → page 69
- [147] M. Jung and P. Hinds. Robots in the wild: A time for more robust theories of human-robot interaction, 2018. → pages 179, 203
- [148] M. F. Jung. Affective grounding in human-robot interaction. In *2017 12th ACM/IEEE Intl Conf on Human-Robot Interaction (HRI)*, pages 263–273. IEEE, 2017. → pages 2, 3, 4, 12, 16, 28, 29, 31, 33, 36, 44, 138, 159, 179
- [149] M. F. Jung, N. Martelaro, and P. J. Hinds. Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 229–236, 2015. → page 204
- [150] M. M. Jung. Towards social touch intelligence: developing a robust system for automatic touch recognition. In *ACM Int’l Conf on Multimodal Interaction (ICMI)*, pages 344–348, 2014. → pages 99, 104, 116, 167
- [151] M. M. Jung, R. Poppe, M. Poel, and D. K. Heylen. Touching the void—introducing cost: corpus of social touch. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 120–127, 2014. → pages 106, 167, 171, 180
- [152] M. M. Jung, X. L. Cang, M. Poel, and K. E. MacLean. Touch challenge’15: Recognizing social touch gestures. In *Proc of the 2015 ACM on Intl Conf on Multimodal Interaction*, pages 387–390, 2015. → pages 13, 14, 65, 99, 116, 180
- [153] M. M. Jung, M. Poel, R. Poppe, and D. K. Heylen. Automatic recognition of touch gestures in the corpus of social touch. *J on multimodal user interfaces*, 11(1):81–96, 2017. → page 117
- [154] H. Kaneko and J. Horie. Breathing movements of the chest and abdominal wall in healthy subjects. *Respiratory care*, 57(9):1442–1451, 2012. → page 80
- [155] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE Trans Pattern Analysis & Machine Intelligence*, 30(12):2067–2083, 2008. → pages 14, 98, 103, 104, 105, 108, 117, 121, 176, 177

- [156] T. Kinnunen and M. Kolehmainen. Touch and affect: Analysing the archive of touch biographies. *Body & Society*, 25(1), 2019. → page 158
- [157] T. A. Klausen, U. Farhadi, E. Vlachos, and J. Jørgensen. Signalling emotions with a breathing soft robot. In *2022 IEEE 5th International Conference on Soft Robotics (RoboSoft)*, pages 194–200. IEEE, 2022. → page 80
- [158] B. C. Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018. → page 13
- [159] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011. → page 180
- [160] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Chiropractic Medicine*, 15(2):155–163, 2016. → page 145
- [161] J. Kortelainen, S. Tiinanen, X. Huang, X. Li, S. Laukka, M. Pietikainen, and T. Seppanen. Multimodal emotion recognition by combining physiological signals and facial expressions: a preliminary study. In *EMBC Annual Conf*, pages 5238–5241, 2012. → pages 98, 105
- [162] G. Kostka, L. Steinacker, and M. Meckel. Between security and convenience: Facial recognition technology in the eyes of citizens in china, germany, the united kingdom, and the united states. *Public Understanding of Science*, 30(6):671–690, 2021. → page 204
- [163] S. D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421, 2010. → page 181
- [164] A. Kron, A. Goldstein, D. H.-J. Lee, K. Gardhouse, and A. K. Anderson. How are you feeling? revisiting the quantification of emotional qualia. *Psychological science*, 24(8):1503–1511, 2013. → page 32
- [165] A. Kron, M. Pilkiw, J. Banaei, A. Goldstein, and A. K. Anderson. Are valence and arousal separable in emotional experience? *Emotion*, 15(1):35, 2015. → page 32
- [166] J. H. Kryklywy, M. R. Ehlers, A. O. Beukers, S. R. Moore, R. M. Todd, and A. K. Anderson. Decomposing neural representational patterns of

discriminatory and hedonic information during somatosensory stimulation. *eneuro*, 10(1), 2023. → page 203

- [167] J. H. Kryklywy, P. Vyas, K. E. Maclean, and R. M. Todd. Characterizing affiliative touch in humans and its role in advancing haptic design. *Annals of the New York Academy of Sciences*, 1528(1):29–41, 2023. → pages 3, 20, 203
- [168] J. S. Kumar and P. Bhuvaneswari. Analysis of electroencephalography (eeg) signals and its categorization—a study. *Procedia engineering*, 38: 2525–2536, 2012. → page 159
- [169] P. Kuppens and P. Verduyn. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26, 2017. → pages 15, 16, 136, 179, 181
- [170] R. Kurzban. The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior*, 25 (4):241–259, 2001. → page 41
- [171] D. Lakens, A. M. Scheel, and P. M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018. → page 149
- [172] G. Lakoff and M. Johnson. *Metaphors we live by*. University of Chicago press, 2008. → pages 25, 26
- [173] S. Lallé, R. Murali, C. Conati, and R. Azevedo. Predicting co-occurring emotions from eye-tracking and interaction data in metatutor. In *International Conference on Artificial Intelligence in Education*, pages 241–254. Springer, 2021. → page 13
- [174] F. Larradet, R. Niewiadomski, G. Barresi, D. G. Caldwell, and L. S. Mattos. Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection. *Frontiers in psychology*, 11:1111, 2020. → page 176
- [175] L. Leahu and P. Sengers. Freaky: performing hybrid human-machine emotion. *Designing Interactive Systems*, pages 607–616, 2014. doi:10.1145/2598510.2600879. URL <http://dl.acm.org/citation.cfm?doid=2598510.2600879>. → pages 28, 31, 32
- [176] J. E. LeDoux. *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster, 1998. → page 13

- [177] M. K. Lee, J. Forlizzi, S. Kiesler, P. Rybski, J. Antanitis, and S. Savetsila. Personalization in hri: A longitudinal field experiment. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 319–326, 2012. → page 205
- [178] M. Leng, Y. Zhao, and Z. Wang. Comparative efficacy of non-pharmacological interventions on agitation in people with dementia: A systematic review and bayesian network meta-analysis. *Int'l J of Nursing Studies*, 102:103489, 2020. → page 158
- [179] R. Levenson. Emotion elicitation with neurological patients. *Handbook of emotion elicitation and assessment*, pages 158–168, 2007. → pages 99, 102, 104, 129, 180
- [180] R. W. Levenson. Autonomic nervous system differences among emotions. *Psychological science*, 3(1):23–27, 1992. → pages 12, 102, 180
- [181] M. D. Lewis. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and brain sciences*, 28(2):169–194, 2005. → page 31
- [182] M. Li and B.-L. Lu. Emotion classification based on gamma-band EEG. In *2009 Annual Int'l Conf of the IEEE Engineering in Medicine and Biology Society*, pages 1223–1226, Sept. 2009. doi:10.1109/IEMBS.2009.5334139. ISSN: 1558-4615. → page 164
- [183] J. Z. Lim, J. Mountstephens, and J. Teo. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors*, 20(8): 2384, 2020. → page 14
- [184] E. W. Lindsey. Relationship context and emotion regulation across the life span. *Emotion*, 20(1):59, 2020. → page 1
- [185] O. V. Lipp and N. Derakshan. Attentional bias to pictures of fear-relevant animals in a dot probe task. *Emotion*, 5(3):365, 2005. → page 80
- [186] Y.-l. Liu, W. Yan, and B. Hu. Resistance to facial recognition payment in china: The influence of privacy-related factors. *Telecommunications Policy*, 45(5):102155, 2021. → page 204
- [187] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005. → page 148

- [188] H.-R. Lv, Z.-L. Lin, W.-J. Yin, and J. Dong. Emotion recognition based on pressure sensor keyboards. In *2008 IEEE Int'l Conf on multimedia and expo*, pages 1089–1092. IEEE, 2008. → pages 9, 156, 159, 171
- [189] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1), 2018. ISSN 20452322. doi:10.1038/s41598-018-32063-4. → page 31
- [190] D. Marino, P. Bucci, O. S. Schneider, and K. 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. → page 32
- [191] F. Martin, K. E. Bachert, L. Snow, H.-W. Tu, J. Belahbib, and S. A. Lyn. Depression, anxiety, and happiness in dog owners and potential dog owners during the covid-19 pandemic in the united states. *PLoS One*, 16(12): e0260676, 2021. → page 79
- [192] K. Matheus, M. Vázquez, and B. Scassellati. A social robot for anxiety reduction via deep breathing. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 89–94. IEEE, 2022. → pages 80, 93, 94
- [193] S. McIntyre, S. C. Hauser, A. Kusztor, R. Boehme, A. Mounbou, P. M. Isager, L. Homman, G. Novembre, S. S. Nagi, A. Israr, et al. The language of social touch is intuitive and quantifiable. *Psychological Science*, 33(9): 1477–1494, 2022. → pages 64, 71
- [194] B. Mesquita, L. F. Barrett, and E. R. Smith. *The mind in context*. Guilford Press, 2010. → pages 136, 159
- [195] W. Miller. Interactive journaling as a clinical tool. *Journal of Mental Health Counseling*, 36(1):31–42, 2014. → page 199
- [196] F. M. Miranda, N. Köhnecke, and B. Y. Renard. Hiclass: a python library for local hierarchical classification compatible with scikit-learn. *arXiv preprint arXiv:2112.06560*, 2021. → pages 164, 168, 169
- [197] A. Montagu. Animadversions on the development of a theory of touch. In *Touch in Early Development*, pages 15–24. Psychology Press, 2014. → pages 1, 41, 203

- [198] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2):119–124, 2013. → pages 15, 136, 141, 159, 160
- [199] T. K. Moriyama, A. Nishi, R. Sakuragi, T. Nakamura, and H. Kajimoto. Development of a wearable haptic device that presents haptics sensation of the finger pad to the forearm. In *2018 IEEE Haptics Symp (HAPTICS)*, pages 180–185. IEEE, 2018. → page 49
- [200] W. Moyle, C. J. Jones, J. E. Murfield, L. Thalib, E. R. Beattie, D. K. Shum, S. T. O’Dwyer, M. C. Mervin, and B. M. Draper. Use of a robotic seal as a therapeutic tool to improve dementia symptoms: a cluster-randomized controlled trial. *Journal of the American Medical Directors Association*, 18(9):766–773, 2017. → page 95
- [201] T. Nakata, T. Sato, and T. Mori. Expression of emotion and intention by robot body movement. *5th Conference on Intelligent Autonomous Systems*, pages 352 – 359, 1998. URL <https://staff.aist.go.jp/toru-nakata/IAS.pdfhttp://apps.isiknowledge.com/full{ }record.do?product=UA{ }search{ }mode=GeneralSearch{ }qid=4{ }SID=3A5Lp5OHF9g93BL3BgD{ }page=1{ }doc=1{ }colname=WOS>. → pages 31, 32
- [202] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Trans Affective Computing*, 6(4):385–394, 2015. → page 105
- [203] A. L. Nichols and J. K. Maner. The good-subject effect: Investigating participant demand characteristics. *The Journal of general psychology*, 135(2):151–166, 2008. → page 129
- [204] Y. Noguchi and F. Tanaka. Omoy: a handheld robotic gadget that shifts its weight to express emotions and intentions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. → pages 5, 10
- [205] M. F. Nolan. Two-point discrimination assessment in the upper limb in young adult men and women. *Physical therapy*, 62(7):965–969, 1982. → pages 45, 49
- [206] D. A. Norman. *Emotional design: Why we love (or hate) everyday things*. Civitas Books, 2004. → page 4

- [207] M. Obrist, S. Subramanian, E. Gatti, B. Long, and T. Carter. Emotions mediated through mid-air haptics. In *Proc of the 33rd Annual ACM Conf on Human Factors in Computing Systems*, pages 2053–2062, 2015. → pages 7, 11, 46
- [208] J. S. Odendaal. Animal-assisted therapy—magic or medicine? *Journal of psychosomatic research*, 49(4):275–280, 2000. → page 98
- [209] A. Öhman and S. Mineka. Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review*, 108(3): 483, 2001. → page 80
- [210] A. Ortony and T. J. Turner. What’s basic about basic emotions? *Psychological review*, 97(3):315, 1990. → page 13
- [211] A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge university press, 1990. → pages 15, 136, 141, 159, 160, 176
- [212] M. A. Otaduy and M. C. Lin. High fidelity haptic rendering. *Synthesis Lectures on Computer Graphics and Animation*, 1(1):1–112, 2006. → page 114
- [213] C. Palestini, M. Minero, S. Cannas, E. Rossi, and D. Frank. Video analysis of dogs with separation-related behaviors. *Applied Animal Behaviour Science*, 124(1-2):61–67, 2010. → page 80
- [214] F. Pallavicini, A. Ferrari, A. Pepe, G. Garcea, A. Zancacchi, and F. Mantovani. Effectiveness of virtual reality survival horror games for the emotional elicitation: Preliminary insights using resident evil 7: Biohazard. In *Int’l Conf on Universal Access in Human-Computer Interaction*, pages 87–101. Springer, 2018. → page 142
- [215] A. Papadopoulou, J. Berry, T. Knight, and R. Picard. Affective sleeve: Wearable materials with haptic action for promoting calmness. In *Distributed, Ambient and Pervasive Interactions: 7th International Conference, DAPI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, pages 304–319. Springer, 2019. → pages 8, 11
- [216] B. Parkinson. Intragroup emotion convergence: Beyond contagion and social appraisal. *Personality and Social Psychology Review*, 24(2): 121–140, 2020. → page 80

- [217] T. Partala and V. Surakka. Pupil size variation as an indication of affective processing. *Int'l J Human-Computer Studies*, 59(1-2):185–198, July 2003. ISSN 1071-5819. → page 104
- [218] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32, pages 8024–8035. Curran Associates, Inc., 2019. URL <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>. → page 164
- [219] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *J of Machine Learning Research*, 12:2825–2830, 2011. → pages 164, 167
- [220] P. Philippot, G. Chapelle, and S. Blairy. Respiratory feedback in the generation of emotion. *Cognition & Emotion*, 16(5):605–627, 2002. → pages 80, 85, 94
- [221] R. W. Picard. *Affective computing*. MIT press, 2000. → pages 12, 14
- [222] D. Playdead. Playdead’s inside. <https://playdead.com/games/inside/>, 2022. Accessed: 2022-04-21. → pages 142, 156
- [223] B. A. Price, R. Kelly, V. Mehta, C. McCormick, H. Ahmed, and O. 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. → page 32
- [224] S. Price, N. Bianchi-Berthouze, C. Jewitt, N. Yiannoutsou, K. Fotopoulou, S. Dajic, J. Virdee, Y. Zhao, D. Atkinson, and F. Brudy. The making of meaning through dyadic haptic affective touch. *ACM Transactions on Computer-Human Interaction*, 29(3):1–42, 2022. → pages 44, 46, 47, 75
- [225] Python. Natural language toolkit - documentation. <https://www.nltk.org/>, 2022. Accessed: 2022-04-21. → page 143
- [226] M. Ragnarsdóttir and E. K. Kristinsdóttir. Breathing movements and breathing patterns among healthy men and women 20–69 years of age: reference values. *Respiration*, 73(1):48–54, 2006. → page 80

- [227] M. M. Rahman, A. K. Sarkar, M. A. Hossain, M. S. Hossain, M. R. Islam, M. B. Hossain, J. M. Quinn, and M. A. Moni. Recognition of human emotions using eeg signals: A review. *Computers in biology and medicine*, 136:104696, 2021. → page 13
- [228] J. Rantala, K. Salminen, R. Raisamo, and V. Surakka. Touch gestures in communicating emotional intention via vibrotactile stimulation. *International Journal of Human-Computer Studies*, 71(6):679–690, 2013. → pages 7, 11
- [229] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2022. URL <https://CRAN.R-project.org/package=psych>. R package version 2.2.3. → page 149
- [230] N. E. Richeson. Effects of animal-assisted therapy on agitated behaviors and social interactions of older adults with dementia. *American Journal of Alzheimer’s Disease and Other Dementias*, 18(6):353–358, 2003. → page 98
- [231] B. Rimé. Interpersonal emotion regulation. *Handbook of emotion regulation*, 1:466–468, 2007. → page 79
- [232] T. Ritchie, J. J. Skowronski, J. Hartnett, B. Wells, and W. R. Walker. The fading affect bias in the context of emotion activation level, mood, and personal theories of emotion change. *Memory*, 17(4):428–444, 2009. → page 160
- [233] C. Rognon, B. Stephens-Fripp, J. Hartcher-O’Brien, B. Rost, and A. Israr. Linking haptic parameters to the emotional space for mediated social touch. *Frontiers in Computer Science*, page 50, 2022. → page 42
- [234] P. V. Rouast, M. T. Adam, and R. Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Trans on Affective Computing*, 12(2):524–543, 2019. → page 160
- [235] G. A. Rousselet. Does filtering preclude us from studying erp time-courses? *Frontiers in psychology*, 3:131, 2012. → pages 159, 164
- [236] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi:10.1037/h0077714. → page 26

- [237] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980. → pages 15, 16, 99, 109, 123, 136, 139, 143, 159, 172, 179
- [238] J. A. Russell. Measures of emotion. In *The measurement of emotions*, pages 83–111. Elsevier, 1989. → page 15
- [239] J. A. Russell, A. Weiss, and G. A. Mendelsohn. Affect grid: a single-item scale of pleasure and arousal. *J Personality & Social Psychology*, 57(3), 1989. → pages 26, 31, 71, 103, 110
- [240] J. Sabourin, B. Mott, and J. C. Lester. Modeling learner affect with theoretically grounded dynamic bayesian networks. In *Affective Computing & Intelligent Interaction*, pages 286–295. Springer, 2011. → page 103
- [241] M. Saerbeck and C. 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. ISBN 9781424448937. doi:10.1145/1734454.1734473. URL <https://www.bartneck.de/publications/2010/perceptionAffectElicitedRobotMotion/saerbeckBartneckHRI2010.pdf><http://portal.acm.org/citation.cfm?doid=1734454.1734473>. → page 31
- [242] E. M. Sahlstein. Relating at a distance: Negotiating being together and being apart in long-distance relationships. *Journal of Social and Personal Relationships*, 21(5):689–710, 2004. → pages 6, 42
- [243] J. Saldien, K. Goris, B. Vanderborght, J. Vanderfaellie, and D. Lefeber. Expressing emotions with the social robot probio. *International Journal of Social Robotics*, 2(4):377–389, 2010. → page 31
- [244] P. Salovey, J. D. Mayer, S. L. Goldman, C. Turvey, and T. P. Palfai. Emotional attention, clarity, and repair: exploring emotional intelligence using the trait meta-mood scale. *Emotion, disclosure, & health*, pages 125–154, 1995. → page 161
- [245] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Symp on Eye Tracking Research & Applications*, pages 71–78. ACM Press, 2000. → page 112
- [246] T. Schneeberger, M. Hladký, A.-K. Thurner, J. Volkert, A. Heimerl, T. Baur, E. André, and P. Gebhard. The deep method: Towards computational modeling of the social emotion shame driven by theory,

introspection, and social signals. *IEEE Transactions on Affective Computing*, 2023. → page 16

- [247] O. Schneider, S. Zhao, and A. Israr. Feelcraft: User-crafted tactile content. In *Haptic Interaction*, pages 253–259. Springer, 2015. → pages 42, 45, 49
- [248] O. S. Schneider, A. Israr, and K. E. MacLean. Tactile animation by direct manipulation of grid displays. In *Proc of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 21–30. ACM, 2015. → pages 42, 45, 49
- [249] O. S. Schneider, H. Seifi, S. Kashani, M. Chun, and K. E. MacLean. Hapturk: crowdsourcing affective ratings of vibrotactile icons. In *Proc of the 2016 CHI Conf on Human Factors in Computing Systems*, pages 3248–3260, 2016. → pages 44, 45
- [250] S. Schneider, D. U. Junghaenel, T. Gutsche, H. W. Mak, and A. A. Stone. Comparability of emotion dynamics derived from ecological momentary assessments, daily diaries, and the day reconstruction method: Observational study. *Journal of Medical Internet Research*, 22(9):e19201, 2020. → page 181
- [251] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. → page 148
- [252] Y. Sefidgar, K. E. MacLean, S. Yohanan, M. Van der Loos, E. A. Croft, and J. Garland. Design and evaluation of a touch-centered calming interaction with a social robot. *Trans Affective Computing*, PP(99):108–121, 2015. → pages 98, 105
- [253] Y. S. Sefidgar, K. E. MacLean, S. Yohanan, H. M. Van der Loos, E. A. Croft, and E. J. Garland. Design and evaluation of a touch-centered calming interaction with a social robot. *IEEE Transactions on Affective Computing*, 7(2):108–121, 2015. → pages 5, 8, 11, 17, 80, 93, 94, 95
- [254] Y. S. Sefidgar, P. Agarwal, and M. Cakmak. Situated tangible robot programming. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 473–482, 2017. → page 83
- [255] H. Seifi and K. MacLean. Exploiting haptic facets: Users’ sensemaking schemas as a path to design and personalization of experience. *Intl Journal of Human-Computer Studies*, 107:38–61, 2017. → page 42

- [256] H. Seifi and K. E. MacLean. A first look at individuals' affective ratings of vibrations. In *2013 World Haptics Conference (WHC)*, pages 605–610. IEEE, 2013. → pages 42, 45, 69
- [257] H. Seifi, K. Zhang, and K. E. MacLean. Vibviz: Organizing, visualizing and navigating vibration libraries. In *2015 IEEE World Haptics Conference (WHC)*, pages 254–259. IEEE, 2015. → pages 44, 45, 69
- [258] A. K. Seth, A. B. Barrett, and L. Barnett. Granger causality analysis in neuroscience and neuroimaging. *Neuroscience*, 35(8):3293–3297, 2015. → page 148
- [259] F. M. Severgnini, J. S. Martinez, H. Z. Tan, and C. M. Reed. Snake effect: A novel haptic illusion. *IEEE Transactions on Haptics*, 14(4):907–913, 2021. → page 45
- [260] J. Seyama and R. S. Nagayama. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence*, 16(4):337–351, 2007. → page 83
- [261] F. Shaffer, R. McCraty, and C. L. Zerr. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in Psychology*, 5, 2014. ISSN 1664-1078. doi:10.3389/fpsyg.2014.01040. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01040>. → page 181
- [262] A. Sharkey and N. Wood. The paro seal robot: demeaning or enabling. In *Proceedings of AISB*, volume 36, page 2014, 2014. → page 5
- [263] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den Broek. Continuous, real-time emotion annotation: A novel joystick-based analysis framework. *Trans on Affective Computing*, 11(1):78–84, 2020. → pages 139, 144
- [264] H. Sharp. *Interaction Design*. John Wiley & Sons, 2003. → page 130
- [265] S. Shen, P. Slovak, and M. 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. → page 31
- [266] G. Sheppes and J. J. Gross. Is timing everything? temporal considerations in emotion regulation. *Personality and Social Psychology Review*, 15(4): 319–331, 2011. → pages 160, 172

- [267] T. Shibata. Ubiquitous surface tactile sensor. In *Robotics and Automation, 2004. TExCRA'04. First IEEE Technical Exhibition Based Conference on*, pages 5–6. IEEE, 2004. → page 14
- [268] T. Shibata and K. Wada. Robot therapy: a new approach for mental healthcare of the elderly—a mini-review. *Gerontology*, 57(4):378–386, 2011. → pages 10, 95
- [269] K. B. Shimoga. Finger force and touch feedback issues in dexterous telemanipulation. In *IEEE Intelligent Robotic Systems for Space Exploration*, pages 159–178, 1992. → page 114
- [270] A. Shojaie and E. B. Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319, 2022. → page 148
- [271] N. J. Shoumy et al. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *J of Network and Computer Applications*, 149:102447, 2020. → pages 158, 160
- [272] C. N. Silla and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22: 31–72, 2011. → page 163
- [273] D. Silvera-Tawil, D. Rye, and M. Velonaki. Interpretation of social touch on an artificial arm covered with an eit-based sensitive skin. *Int'l J Social Robotics*, 6(4):489–505, 2014. → pages 107, 156
- [274] A. Simoës-Perlant, C. Lemerrier, C. Pêcher, and S. Benintendi-Medjaoued. Mood self-assessment in children from the age of 7. *Europe's Journal of Psychology*, 14(3):599, 2018. → page 139
- [275] J. Smith and K. MacLean. Communicating emotion through a haptic link: Design space and methodology. *International Journal of Human-Computer Studies*, 65(4):376–387, 2007. → pages 7, 11
- [276] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2): 211–223, 2011. → page 14
- [277] S. Song and S. 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. → page 31

- [278] S. H. Sperry, M. A. Walsh, and T. R. Kwapil. Emotion dynamics concurrently and prospectively predict mood psychopathology. *Journal of affective disorders*, 261:67–75, 2020. → pages 16, 141, 151
- [279] T. Staudigl, S. Hanslmayr, and K.-H. T. Bäuml. Theta oscillations reflect the dynamics of interference in episodic memory retrieval. *Journal of Neuroscience*, 30(34):11356–11362, 2010. → pages 171, 172
- [280] W. D. Stiehl, C. Breazeal, K.-H. Han, J. Lieberman, L. Lalla, A. Maymin, J. Salinas, D. Fuentes, R. Toscano, C. H. Tong, et al. The huggable: a therapeutic robotic companion for relational, affective touch. In *ACM SIGGRAPH 2006 emerging technologies*, pages 15–es. ACM, 2006. → pages 95, 97
- [281] W. D. Stiehl, J. K. Lee, C. Breazeal, M. Nalin, A. Morandi, and A. Sanna. The huggable: a platform for research in robotic companions for pediatric care. In *Proceedings of the 8th International Conference on interaction Design and Children*, pages 317–320. ACM, 2009. → page 4
- [282] J. N. Stinson. Improving the assessment of pediatric chronic pain: harnessing the potential of electronic diaries. *Pain Research and Management*, 14(1):59–64, 2009. → page 32
- [283] S. Strohkorb Sebo, M. Traeger, M. Jung, and B. Scassellati. 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. → page 31
- [284] R. Strong, B. Gaver, et al. Feather, scent and shaker: supporting simple intimacy. In *Proceedings of CSCW*, volume 96, pages 29–30, 1996. → page 44
- [285] G. M. Sullivan and A. R. Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013. → page 32
- [286] D. Sun, P. Paredes, and J. Canny. Moustress: detecting stress from mouse motion. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–70, 2014. → pages 9, 11

- [287] J. T. Suvilehto, E. Glerean, R. I. Dunbar, R. Hari, and L. Nummenmaa. Topography of social touching depends on emotional bonds between humans. *Proceedings of the National Academy of Sciences*, 112(45): 13811–13816, 2015. → page 83
- [288] V.-C. Ta, W. Johal, M. Portaz, E. Castelli, and D. Vaufreydaz. The grenoble system for the social touch challenge at icmi '15. In *Proc 2015 ACM on Int'l Conf on Multimodal Interaction*, pages 391–398, 2015. → page 99
- [289] D. Tam, K. E. MacLean, J. McGrenere, and K. J. Kuchenbecker. The design and field observation of a haptic notification system for timing awareness during oral presentations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1689–1698, 2013. → page 11
- [290] H. Tan, A. Lim, and R. Traylor. A psychophysical study of sensory saltation with an open response paradigm. In *the 9th Ann. Symp. on Haptic Interfaces for Virtual Environment and Teleoperator Systems, ASME/IMECE*, volume Vol. 69-2, pages 1109–1115. S. S. Nair, 2000. → page 45
- [291] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak. Eye-tracking analysis for emotion recognition. *Computational intelligence and neuroscience*, 2020, 2020. → page 14
- [292] T. Terkildsen and G. Makransky. Measuring presence in video games: An investigation of the potential use of physiological measures as indicators of presence. *Int'l Journal of Human-Computer Studies*, 126:64–80, 2019. → page 142
- [293] Y. Terzioğlu, B. Mutlu, and E. Şahin. Designing social cues for collaborative robots: the role of gaze and breathing in human-robot collaboration. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 343–357, 2020. → page 80
- [294] M. Teyssier, G. Bailly, C. Pelachaud, and E. Lecolinet. Conveying emotions through device-initiated touch. *IEEE Trans on Affective Computing*, 2020. → pages 1, 42
- [295] T. Thanapattheerakul et al. Emotion in a century: A review of emotion recognition. In *Int'l Conf on advances in information technology*, 2018. → pages 158, 160

- [296] E. H. Thompson and J. A. Hampton. The effect of relationship status on communicating emotions through touch. *Cognition and Emotion*, 25(2): 295–306, 2011. → pages 42, 47, 69
- [297] E. Tieppo, R. R. d. Santos, J. P. Barddal, and J. C. Nievola. Hierarchical classification of data streams: a systematic literature review. *Artificial Intelligence Review*, pages 1–40, 2022. → page 163
- [298] J. B. Torre and M. D. Lieberman. Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review*, 10(2):116–124, 2018. → page 160
- [299] P. M. Ullrich and S. K. Lutgendorf. Journaling about stressful events: Effects of cognitive processing and emotional expression. *Annals of Behavioral Medicine*, 24(3):244–250, 2002. → page 199
- [300] G. A. Van Kleef and S. Côté. The social effects of emotions. *Annual review of psychology*, 73:629–658, 2022. → pages 79, 80
- [301] B. Vanderborght, R. Simut, J. Saldien, C. Pop, A. S. Rusu, S. Pintea, D. Lefeber, and D. O. David. Using the social robot probio as a social story telling agent for children with asd. *Interaction Studies*, 13(3):348–372, 2012. → page 95
- [302] P. Verduyn. Emotion duration. In *Affect Dynamics*, pages 3–18. Springer, 2021. → pages 159, 171
- [303] P. Verduyn, I. Van Mechelen, and F. Tuerlinckx. The relation between event processing and the duration of emotional experience. *Emotion*, 11(1): 20, 2011. → page 164
- [304] P. Verduyn, P. Delaveau, J.-Y. Rotgé, P. Fossati, and I. Van Mechelen. Determinants of emotion duration and underlying psychological and neural mechanisms. *Emotion Review*, 7(4):330–335, 2015. → pages 159, 160, 164
- [305] K. Wada and T. Shibata. Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. *IEEE Trans on Robotics*, 23(5):972–980, 2007. → page 97
- [306] K. Wada, Y. Ikeda, K. Inoue, and R. Uehara. Development and preliminary evaluation of a caregiver’s manual for robot therapy using the therapeutic seal robot paro. In *ROMAN*, pages 533–538, 2010. → page 4

- [307] M. Wagner, Y. Sahar, T. Elbaum, A. Botzer, and E. Berliner. Grip force as a measure of stress in aviation. *The International Journal of Aviation Psychology*, 25(3-4):157–170, 2015. → page 13
- [308] X.-W. Wang, D. Nie, and B.-L. Lu. Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129:94–106, Apr. 2014. ISSN 09252312. doi:10.1016/j.neucom.2013.06.046. URL <https://linkinghub.elsevier.com/retrieve/pii/S0925231213009867>. → pages 164, 171
- [309] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988. → pages 103, 136, 139, 143, 159
- [310] K. Westlund, M. Jacqueline, S. Jeong, H. W. Park, S. Ronfard, A. Adhikari, P. L. Harris, D. DeSteno, and C. L. Breazeal. Flat vs. expressive storytelling: Young children’s learning and retention of a social robot’s narrative. *Frontiers in human neuroscience*, 11:295, 2017. → page 32
- [311] C. J. Willemse, G. M. Munters, J. B. van Erp, and D. Heylen. Nakama: A companion for non-verbal affective communication. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 377–378, 2015. → pages 7, 11
- [312] C. J. Willemse, D. K. Heylen, and J. B. van Erp. Communication via warm haptic interfaces does not increase social warmth. *Journal on multimodal user interfaces*, 12:329–344, 2018. → page 11
- [313] N. Williams, K. MacLean, L. Guan, J. P. Collet, and L. Holsti. Pilot testing a robot for reducing pain in hospitalized preterm infants. *OTJR: occupation, participation and health*, 39(2):108–115, 2019. → page 93
- [314] T. Williams, D. Thames, J. Novakoff, and M. 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. → page 31
- [315] I. H. Witten. Data mining with WEKA. *UWaikato, New Zealand*, 2013. → page 122

- [316] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proc of the SIGCHI Conf on human factors in computing systems*, pages 143–146, 2011. → page 169
- [317] T. Xie, M. Cao, and Z. Pan. Applying self-assessment manikin (sam) to evaluate the affective arousal effects of vr games. In *Proc in Int’l Conf on Image & Graphics Processing*, pages 134–138, 2020. → page 139
- [318] T. Xue, S. Ghosh, G. Ding, A. El Ali, and P. Cesar. Designing real-time, continuous emotion annotation techniques for 360 vr videos. In *Extended Abstracts of Int’l Conf on Human Factors in Computing Systems*, pages 1–9, 2020. → pages 139, 144
- [319] S. Yohanan and K. E. MacLean. The haptic creature project: Social human-robot interaction through affective touch. In *Proceedings of the AISB 2008 Symposium on the Reign of Catz & Dogs: The Second AISB Symposium on the Role of Virtual Creatures in a Computerised Society*, volume 1, pages 7–11. Citeseer, 2008. → pages 5, 8, 11, 14
- [320] S. Yohanan and K. E. MacLean. Design and assessment of the haptic creature’s affect display. In *ACM/IEEE Int’l Conf on Human-Robot Interaction (HRI ‘11)*, pages 473–480, Lausanne, SW, 2011. → pages 23, 31
- [321] S. Yohanan and K. E. MacLean. The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *International Journal of Social Robotics*, 4(2):163–180, 2012. → pages 5, 12, 97, 99, 104, 156, 158, 202
- [322] N. Yoshida and T. Yonezawa. Investigating breathing expression of a stuffed-toy robot based on body-emotion model. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 139–144, 2016. → page 83
- [323] J. Zaki and W. C. Williams. Interpersonal emotion regulation. *Emotion*, 13(5):803, 2013. → page 79
- [324] B. Zhang, G. Essl, and E. Mower Provost. Automatic recognition of self-reported and perceived emotion: Does joint modeling help? In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 217–224, 2016. → page 208

- [325] L. H. Zhang, P. Bucci, X. L. Cang, and K. MacLean. Infusing cuddlebits with emotion: Build your own and tell us about it. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2018. → page 206
- [326] W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Trans on autonomous mental development*, 7(3), 2015. → pages 162, 171
- [327] Y. Zhou, A. Murata, and J. Watanabe. The calming effect of heartbeat vibration. In *2020 IEEE Haptics Symposium (HAPTICS)*, pages 677–683. IEEE, 2020. → pages 8, 11