The Computation of Meaning:

From Embodied Emotions to Cognitive Schemas

by

Paul Bucci

B.A., The University of British Columbia, 2012
B.Sc., The University of British Columbia, 2015
M.Sc., The University of British Columbia, 2017

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)

October 2024

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

The Computation of Meaning:
From Embodied Emotions to Cognitive Schemas

submitted by Paul Bucci in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in Computer Science

Examining Committee:

Ivan Beschastnikh, Associate Professor, Computer Science, UBC
Supervisor

Leanne Currie, Associate Professor, Nursing, UBC
Supervisory Committee Member

Rachel Pottinger, Professor, Computer Science, UBC
University Examiner

Julia Bullard, Assistant Professor, Information School, UBC
University Examiner

Malte Jung, Associate Professor, Information Science, Cornell University
External Examiner

Additional Supervisory Committee Members:

Tamara Munzner, Professor, Computer Science, UBC
Supervisory Committee Member

Abstract

How do we compute meaning? To make something computable, we must reduce the world to logical operations on electrical signals. However, our human experience is that the world has an uncomputable, meaningful aspect that seems to defy mere information processing. The quantitative world of computing demands measurable, objective signals to be translated into the qualitative world of affect, emotion, and meaning. Is it possible to make the two worlds of qualitative and quantitative meet?

In this dissertation, I report on, analyze, and draw conclusions from two multi-part projects that attempt to answer this question from different perspectives using interactive systems and machine learning. First, we look at computing meaning by attempting to detect emotions using signals derived from the body such as heart rate, brain waves, and gestures. Then, we look at computing meaning by making connections between documents to support thematic exploration of large document corpora.

My contributions in this dissertation are:

- A critical theoretical and methodological proposition for computationally representing, sensing and displaying real-time emotions.
- A synthesis of the theoretical and pragmatic basis of therapeutic care methods and their meaning for affective robotics, with an accompanying account of the constructed nature of emotions for HRI applications.
- The design and evaluation of a system (called Teleoscope) for capturing underlying meaning in documents through interaction with machine learning systems.
- An extension to thematic analysis for data curation to create meaning in large text datasets which we call thematic exploration, and a methodological concept of schema crystallization.

Through these projects, an underlying understanding of meaning-making as an embedded, embodied, emergent, interactive phenomenon is articulated. That is to say, meaning is *embedded* in a culture and environment,

embodied in the whole of a person, and *emerges* through the process of interaction between a person, themselves, other people, and their environment. By understanding these epiphenomenal interactions, designers may be enabled to create computational systems that facilitate richer meaning-making.

Lay Summary

Computer Scientists often make artificial intelligence (AI) models that try to detect meaningful things like emotions, or perform meaningful procedures like categorizing texts. This dissertation reports on two lines of research: one where we tried to detect people's emotions using sensors on the body, and another where we made a system for exploring millions of documents to find meaningful themes (called Teleoscope). The system for detecting emotions did not work in a way that I liked, and so I analyzed why that might be and critiqued common understandings of detecting emotions in Computer Science. Then, when we built Teleoscope, we followed a careful process of working with our end users, and we were much more satisfied with the results. As such, I analyzed why I think Teleoscope works well, and what might be going on when people try to use systems like Teleoscope to discover meaningful themes in documents.

Preface

The research presented in this dissertation is the result of many collaborations, publications, and required multiple UBC Research Ethics Board approvals.

Included in the text of the dissertation are only papers that I was the first author on and claim the primary contribution for, but there are analyses and references to other papers that I also contributed heavily to and report on.

Chapter 3, includes the text of *Real Emotions Don't Stand Still*, which was published in the 2019 8th International Conference on Affective Computing and Intelligent Interaction, and co-authored by Xi Laura Cang, Hailey Mah, Laura Rodgers, Karon MacLean:

- Bucci, P., Cang, X. L., Mah, H., Rodgers, L., & MacLean, K. E. (2019). Real emotions don't stand still: Toward ecologically viable representation of affective interaction. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-7). IEEE.

My contribution for this paper is the majority of the conceptual formulation, writing, and analysis, but the paper was the result of a long-term collaboration between Laura Cang and myself while supervised by Karon MacLean. Together, we supervised a large number of Research Assistants and collaborated with people from other departments. Papers that I make reference to and perform a meta-analysis of but do not claim as a contribution for this dissertation, as well as the papers which were reported on for my Master's Degree and therefore are not part of this dissertation's contribution are available in Appendix A.

Our undergraduate RAs included: Hafsa Zahid, Andrew Moore, Liz Koswara-Simms, Gabby Savage, Liam Butcher, Sherry Yuan, Eileen Ong, QiQi Li, Hannah Elbaggari, Linda Jiang, Sean Fernandes, Anushka Agrewal, Anita Shah, Hailey Mah, Drishtti Rawat, Qianqian Feng, Zefan Sramek, Laura Rodgers, Minjia Zhan, Tyler Malloy, Aiden Smith, Bryan Lee, Mario

Cimet. Each was responsible for a one- to two-term project where they either built a component of our sensing and robotic systems, ran participants, or performed data analysis. All conceptual and project management was performed by myself or Laura Cang.

Chapter 4, *Affective robots need therapy* is a summation of my learning from the above and includes entirely new recommendations and analyses that are largely independent of the work with Laura Cang and Karon MacLean. Except for a couple of paragraphs, this work was almost entirely written by me, and edited by my co-authors David Marino and my supervisor Ivan Beschastnikh. It includes the text from the following publication:

- Bucci, P., Marino, D., & Beschastnikh, I. (2023). Affective robots need therapy. ACM Transactions on Human-Robot Interaction, 12(2), 1-22.

Chapter 5, *Teleoscope Systems Paper*, is the culmination of three years of coding and design work on Teleoscope. I was the principal systems architect, designer, and coder, however, I again oversaw a large number of undergraduate Research Associates who participated in one- to two-term projects coding and designing a part of the system: Armin Talaie, Aanandi Sidharth, Qiyu Zhou, Kenny Averna, Dhruv Khanna, Patrick Lee, Sol Lee, Crystal Lee, Leo Foord-Kelsey, Alamjeet Singh, Prayus Shrestha, Florentina Simlinger, Vita Chan. I was the sole writer of the paper and my supervisor Ivan Beschastnikh was the primary editor.

- Bucci, P., Foord-Kelcey, L., Lee, P. Y. K., Singh, A., & Beschastnikh, I. (2024). Crystallizing Schemas with Teleoscope: Thematic Curation of Large Text Corpora. arXiv preprint arXiv:2402.06124v2.

Chapter 6, *Crystallizing Schemas through Thematic Externalization with Large Corpora*, is the summary of my and Ivan Beschastnikh's process of using Teleoscope. I was the sole writer and my supervisor Ivan Beschastnikh was the editor.

We are now in the process of incorporating a business that will run a version of Teleoscope. That is being undertaken and lead by myself, but includes Leo Foord-Kelsey, Alamjeet Singh, and Patrick Lee, as well as a non-student collaborator, Nathaniel Ki.

All research was conducted under the review of UBC's Behavioural Research Ethics Board under the following approvals:

- Interactive Affective Touch: H15-02611
- Conflicting Identities: H21-00285

- Teleoscope: H22-03775
- RiCC:Childcare: H19-00482
- TAMER: H09-02860
- Touch sensing interactions: H16-01549

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgements

This dissertation would not have come to fruition without the influence and help of a great number of people. Since it is not often that I have a formal way to reflect and thank people, I will make full use of the chance here.

First, I must thank my supervisor Dr. Ivan Beschastnikh for taking a chance on me and providing consistent, kind support. I have learned a lot just from watching your approach to mentorship and work, and have greatly appreciated your writing and technical advice.

Second, my committee members, Drs. Leanne Currie and Tamara Munzner. Leanne, I have appreciated your consistent kind help and being able to work with you in a variety of capacities over the years. Tamara, I have seen you sit down with many people (including me) and help them clarify their research. I have often tried to emulate your clarity and patience. Also your approach to mentorship, which I have tried to emulate with all of my students (including requiring them to meet with me during my classes, taken from your vis class!).

Third, the examining committee, including Drs. Rachel Pottinger, Julia Bullard, and Malte Jung. It was gratifying to have people I know and respect read my work so closely and well. Rachel, it was particularly meaningful to me that you were able to see me through the beginning and end of my journey through CS at UBC.

Fourth, my collaborators in Systopia and SPIN. Particularly Drs. Oliver Schneider and Hasti Seifi for mentoring me early on, Dr. Margo Seltzer for mentorship later on, and Dr. Laura Cang for the years of long-term collaboration and friendship. Also the long list of collaborators without whom none of this work would have been possible: Aanandi Sidharth, Qiyu Zhou, Kenny Averna, Dhruv Khanna, Patrick Lee, Sol Lee, Crystal Lee, Leo Foord-Kelsey, Alamjeet Singh, Prayus Shrestha, Florentina Simlinger, Vita Chan, Dr. Merel Jung, Dr. Jussi Rantala, Hanieh Shakeri, Dilan Ustek, Kevin Chow, Dr. Soheil Kianzad, Hafsa Zahid, Andrew Moore, Liz Koswara-Simms, Gabby Savage, Liam Butcher, Sherry Yuan, Eileen Ong, QiQi Li, Hannah Elbaggari, Linda Jiang, Sean Fernandes, Anushka Agrewal, Anita Shah, Hailey Mah, Drishtti Rawat, Qianqian Feng, Zefan

Sramek, Laura Rodgers, Minjia Zhan, Tyler Malloy, Aiden Smith, Bryan Lee, Mario Cimet, Lotus Zhang, Sophia Chen, Anasazi Valair, Lucia Tseng, David Marino, Yana Pertels (RIP), and Alicia Woodside, and others that accidentally didn't make it into the lists I've compiled over the years.

Fifth, the academic friends and mentors I had, particularly Dr. Eric Vatikiotis-Bateson (RIP), Dr. Chris Mole, Christine D'Onofrio, Dana Claxton, Richard Prince, Kevin Murphy, and Dr. Bob Pritchard, the ThunderBots@Home team, Carson Logan, Dr. Madision Elliot, Lisa Shiozaki, Rama and the VisCog coding team. I wouldn't have done anything remotely related to computation and design if you folks hadn't pushed me towards visualization, interactive art and cognitive systems.

Sixth, the friends and mentors I've had who taught me how to be an artist, writer, and designer. None of this could have happened without the things I learned gratefully or ungratefully from you all. There are far too many names to list, but if you search through Ubyssey mastheads for about six or so years from 2006 on, each person is someone I've known far too well as a personal friend and somehow both exacerbated and smoothed out my rough edges as a writer, manager, collaborator and person. Champagne Choquer, Oker Chen, Goh Iromoto, Gerald Deo, and Kate Barbaria for teaching me how to be a designer over many gruelling hours and late nights. Goh and Kate particularly for shaping my whole designer's mind. Rico Moran and Goh again for teaching me film and business. The Syrup Trap, particularly David (again) and Nick Zarzycki, but also Jimmy, Nathan, Matthew, Bryce, Winnie, and more; Balkan Haus boys Nick and David again but also Jonny, Chris, and more. EyeMole (David again, but also Vesta Sahatçiu and Yana again); the so-called Shifting Collective (especially Rhys Edwards, Pauline Petit, Adrian Diaz, and Andy Keech); and Blind Tiger (David again??).

Last, the people who have brought light into my life as I rewrite it. David and Nick again and always. Lily Ivanova who has really been on the journey with me in so many ways it's hard to even imagine (how many more PhDs in life can we get?). Fr. Neil, Mama SheShe, and Daniel Amy who brought me a poetic calm while collecting my thoughts for this thing. The whole St. John's crew supporting me, particularly the Hewletts, John, Carl, Wiley, and Laurence. Jesse Amy as we move through the unfolding of our seasons together. The Abby Crew, including Jared, Joe, Tamara, Cam, Dylan, and Mike. Mackie and my family. Those who choose to stay anonymous and those I have left out of the list but not forgotten.

Funding included NSERC Discovery Grant RGPIN-2020-05203, the DFP Project Stimuls Grant 2022, UBC BCGS, NSERC CGSD.

Dedication

This dissertation is dedicated to mentors and friends who have passed on. Each made my life difficult and meaningful in their own way:

Fernie, who protected my first playground.
Eric, who refused to define cognitive systems.
Yana, may you find peace.

Chapter 1

Introduction

It came to be that I[1] was working at a children's hospice in the middle of my dissertation. Ostensibly, my job was to observe how clinical staff and patients were interacting via emotional touch. My lab had been working on a fabric touch sensor for a social robot, and I had the job of understanding how these could be used in a therapeutic setting. Being someone who was comfortable making physical things due to my background in visual arts and design, my research had focused on making small, furry robots that were wrapped in this touch sensor. With my lab mates, I also worked on related design and engineering, that is, training machine learning (ML) models of social touch, programming robot behaviour, and figuring out how and why these robots could be used.

During one of my first projects, we trained an ML model to detect touch gestures such as pat, tickle, scratch, etc., while the sensor was draped over a robot. To our surprise (perhaps unwarranted), we discovered that we could differentiate both participants and touch gestures, i.e., detect that your *pat* or *tickle* is different than my *pat* or *tickle*. This led to the question: what if we could determine features of one participant's touch at different times? And different times might mean different moods. Maybe people scratch their pets differently based on how they are feeling. Perhaps *how* somebody touched may provide insight into their emotional state.

The chain of logic continued: if a robot could detect someone's emotional state, maybe it could provide a meaningful therapeutic interaction. A robot that enacted sympathy. A robot that reacted with care and provided therapeutic touch. Knowing that this could be a quagmire of confounding factors, we pared the project down to be a robot with one motor and no facial features: just a fluffball that made breathing motions. After two years of my master's work designing the robot's body and behaviours, we were ready to take the robot into a clinical setting. We contacted a dementia ward where they were already using furry robots and a children's hospice, and met with

[1] In this dissertation, I will use "I" when referring to my own experience or singular work, but for instances where I was part of a team (e.g., manuscripts), I will use "we" to refer to myself and my team.

clinical staff in both. I headed the hospice work, and my lab partner headed the dementia work.

My experience in the hospice profoundly changed me. Or it may have simply helped me concretely realize notions that I had suspected to be true. I entered their volunteer training program and started observations. Immediately, I realized that there was no possible way to build a robot that could, on its own, meaningfully interact with people who were preparing for death (either their own or their loved ones). The dream of therapeutic touch driven by the robot was gone. My lab partner and I had suspected that we might eventually learn that the robot was not useful, but it was totally confirmed at a gut level as soon as we walked in and listened for a day. However, we also saw very clear and beautiful opportunities for meaningful technologically-mediated interactions. In fact, the hospice already had many interactive devices, practices, and technologies that they were using, which I will mention below. We just needed to understand the hospice environment itself to see where we might fit in.

The shift in perspective was both monumental and subtle. We had been approaching our robot design as if we could—and should—detect emotional states as part of a system of meaning. This assumes that the robot has an internal model of the interactor's emotional state, a machine model approximating the human ability to guess how other people are feeling. It also assumes that the robot can make appropriate emotional decisions based on this model. These are significant assumptions.

However, the technology in use at the hospice was both symbolically and practically meaningful. Not because of something *within* the system, but instead *outside* the system, in the context, use, and playful interactions of the human therapists who worked with the technology. Good technology design here meant designing devices that were simple, easy-to-use, and facilitated self-knowledge and play. Devices would be more like improvisational instruments with rather than providing a pre-determined interaction. Sort of like the difference between a video game and a lego set: the video game must be fully immersively programmed to be impactful, whereas a lego set facilitates endless improvisation by recombining simple modules.

The standard approach to touch-based interaction at the hospice was an improvisational communication technique called *intensive interaction*[2]. For non-verbal, wheelchair-bound children, it is one of the only ways to sustain profound, meaningful communication. An interactor ensures that they

[2]Intensive interaction is well-documented on their website at intensiveinteraction.org; watching a video helps to clarify the communicative potential.

are eye-level with the child, and watches their face carefully for emotional signals. This may be difficult, because the children do not have consistent control of their muscles. They hold the child's hand, or whatever part of the body may be expressive in the moment. The interactor speaks their intentions out loud and reinforces with their entire body, e.g., drawing the child's attention to a shiny pillow by bringing the pillow close, saying "Look at the shiny pillow" repeatedly, looking at it themselves, and bringing the pillow into physical contact with the child. The interactor will describe the child's experience using repetitive sensation and emotion words, attending to any sign of recognition. If the child makes a sound, the interactor will repeat it, guessing at the meaning. To an observer, it sounds quite silly—and it should! Mutual joy is part of the effectiveness of the technique. With enough time and attention, the interaction develops into a sort of improvised, joyful, touch-and-silly-sound-based "language."

Clearly, if you watch them, the child has a meaningful experience. Due to the fact that they are non-verbal, it is difficult to know exactly what their experience is in words, but the joy on their faces is unmistakable. The interactors will use this technique with any being, object, genre, material or device that is available. For example, a pet rabbit was occasionally brought in to the hospice, and the children would touch the rabbit with a therapist using intensive interaction. Or the music therapist would use this technique to improvise songs. Or, the art therapist might do this with paint and colour. The possibilities were endless.

In terms of interactive technologies, a few design opportunities stood out. They had a *Snoezelen Room*[3], which is a multi-media interactive sensory stimulation chamber that can be soothing to certain children. They used a variety of vibrating and texture-based toys to stimulate touch experiences. And they would use a variety of digital music systems combined with the above to create improvisational multi-media experiences that could be seen, heard, and felt at the same time.

Unfortunately, the COVID-19 pandemic cut short my observations at the hospice. I would love to have had the rest of my dissertation be about the technologies I designed for the hospice, since I had many profound experiences there. It was not to be. However, the experiences I did have drove home deep experiential realities of the "meaning of meaning," the hubris of certain approaches to affective computing (including my own, previously), and clear underlying sense of where digital design can support meaning in human interactions rather than overtaking it. It made me want to try to

[3] snoezelen.info

understand what "meaning" could mean computationally.

Meaning at the hospice was multifaceted. Paradoxically, the constant presence of death created an atmosphere of joy rather than sadness. However, the reality of death needed to be processed. The hospice community developed symbols such as lighting a green lamp when a child was close to death, then turning it off when the child had died. This way, staff and visitors could know at a glance whether it was a time of mourning. Candles were lit for the departed. Paintings that the children made were kept. Meals were eaten communally. And most importantly, all children were treated like children. They were given education, whatever that might mean for them. Why teach a dying child? Because that is what people do with children.

I learned there that emotions are part of a process of experience, not a state. External objects are symbols which are imbued with meaning through practice, rather than having intrinsic meaning. People need external symbols to make sense of their internal being, which is fluctuating, contingent, and multi-faceted. Even profoundly traumatic experiences like death can be reckoned with meaningfully if the meaning is actively created through and with externalized symbols. However, symbols don't have meaning on their own: they need to be created, curated, and reinforced by a whole community.

1.1 Defining meaning

Theories about symbols and meaning have been a large part of the 20th century philosophical, psychological, and sociological tradition. De Saussure famously set up the dichotomy of the signifier (the symbol) and signified (what is being symbolized) (West [2005]). The arbitrary nature of language symbols is a foundational feature of linguistics, and refers to the fact that, most often, symbols can have features that have no apparent direct connection to an objective feature. Most words do not sound like the things they represent: nothing about the word 'cat' indicates a furry animal with pointy ears. Only in rare cases such as onomatopoeia does a word 'sound like' anything it refers to (e.g., 'meow' sounds a bit like a cat's vocalizations).

Since De Saussure, many empirical works have been done on language acquisition and cultural production of meaning. Elizabeth Bates presents a theoretical synthesis of work on the childhood development of language (Bates [2014]), wherein she takes the idea of the arbitrariness of language, breaks it down, and situates it in observations of children's behaviour. An example she gives is the word 'shoe': a child does not 'know' anything about the larger cultural meaning of shoes, nor the importance of putting them

on, nor particularly why the word 'shoe' is attached to a whole system of action, but they learn slowly through repetition and shared attention that the symbol is associated with an object and practice. Michael Tomasello further articulates the ability of children to develop a theory of mind, that is, an understanding of how another thinks, to share attention on things in the world and create meaning as a result (Tomasello [2014]).

However, there are two senses of the word 'meaning' at play here. As opposed to the symbol-object referent paradigm above, Gilad Hirschberger discusses how meaning develops in his work 'Collective Trauma and the Social Construction of Meaning' (Hirschberger [2018]). In this sense of the word, traumatic events are given meaning through a collective response to identity threats. This sense of meaning is more about the narrative around an event which is often condensed into a symbol, but instead refers to larger aesthetic, moral, and affective experiences rather than a clearly-defined object or action.

The way that the two senses of the word meet are in social constructivist theories of meaning. Sociologists Berger and Luckmann are *the* foundational thinkers in asserting that our experience of reality is constructed through social forces (Berger and Luckmann [2016], Pfadenhaueris and Knoblauch [2019]). Jean Mandler distills a general concept of psychological *schemas* into a wide-ranging theory of different schema types that brings together meaning-as-narrative and and meaning-as-object-referent (Mandler [2014]).

In this dissertation, I will take both a poetic and scientific approach to meaning. In the poetic sense, meaning is accumulative, interpretive, shifting, and rich with associations. In the scientific sense, meaning is definitive, bounded, and more useful when stable and precise. A unifying feature of both is that *meaning is understood as relationship*, whether that is between symbols, concepts, or real, physical objects. Philosopher Hilary Putnam discusses meaning at length in his work *The Meaning of "Meaning"* (Putnam [1975]), questioning the largely semiotic understanding taken by De Saussure with regards to the relationship between signified and signifier as an ontological relation, that is, how does the really-existing thing relate to the way we represent it? Putnam talks about water (the cultural concept, word, or signifier) vs. the substance that is present (the signified) vs. H2O vs. a hypothetical molecule that functions exactly like water but does not have the same chemical formula.

If the word "water" refers to a substance that we can use to drink, put out fires, freeze into ice, etc., then the chemical formula would not matter to any human linguistic community (other than chemists), and the meaning would be practically identical to the meaning of H2O. That this

might be unsatisfying to those of us who "know" that water is chemically H2O is mostly irrelevant. I, personally, keep thinking that of course the "meaning" of the water-stand-in-substance would change when subject to the right atomic experiment, but that is exactly the point: except for a rare few with access to equipment complicated enough to test it, there would be no *meaningful* difference. And even after the experiment is over, the scientist who "knows" the difference would only know the difference insofar as he set the "fake" water apart from the "real" water; mixing together would eradicate the *meaningful* difference immediately. The *meaning* is constructed culturally, and, I would argue, poetically.

A good example of this poetic construction is "holy water." Orthodox Christian theologian Fr. Alexander Schmemann addresses the meaning of holy water exactly as something that is set apart, treated differently than "normal water," but is clearly chemically identical to water (Schmemann [1973]). He takes a teleological stance, that is, addressing the *purpose* of water, and saying that holy water most fully embodies what it *means* to be water. Not only does it clearly fulfill all of the practical realities of water, but because there has been human and divine effort involved in setting it apart, it takes on mystical qualities beyond water. In a water-blessing ceremony, a whole church community will come together for hours of prayer, putting the water in a special container, and invoking ancient rites and interventions of the divine creator. Even if a scientist might scoff at the water now having different physical properties, the point is that a community of people treats it differently, therefore it clearly takes on a different *meaning* than "normal water."

When I refer to *making* meaning, it may not take the effort of a priest and a large group of people, but it serves as a strong example of what might be involved in the actual *making* of meaning. Elements of authority, tradition, process, community, liturgical practice, belief and personal aesthetic sense are involved in academic qualitative research—perhaps not surprisingly, due to the origin of hermeneutic analysis in establishing definitive biblical interpretation. Quantitative research has its own versions of meaning-making through scientists (priests) performing experiments (liturgical practice) and adding to the literature (tradition), with the added benefit of establishing (holy) causality.

For the purpose of this dissertation, I will invite the reader to meditate on the intentionally pithy statement that *meaning is relationship*. Read the statement poetically to capture all senses of the words *meaning* and *relationship*. There is a strong computational reality to the statement. For

computer scientists, meaning can be quite literally constructed as data[4] objects that have pointers to other data objects. A graph-theoretic understanding of meaning could be operationalized as the measurable extent to which a vertex is connected to other vertices. In NLP, explicit databases of word senses called "ontologies" are sometimes constructed as a tree, and the meaning of a word is procedurally defined as a lookup distance from one word to another word. For example, a program might assign a probability score to an instance of the word *dog* in a sentence as *animal* due to the short parent-child distance in the ontology. Or, similarly, the word *cat* has a meaning that is close to *dog* due to the sibling relationship to *animal*.

More up-to-date NLP systems will use word distribution and co-location to construct multi-dimensional vector representations of words called "embeddings" which calculate semantic similarity as the distance between vectors. This is a statistical approach, where meaning is entirely derived from real-world use cases. *Dog* is semantically similar to *cat* because the data may have many instances where either might show up near the word *pet*, not because of a scientific, taxonomic relationship that the computer has been programmed with. "Meaning" is what is accessible in a high-dimensional volume near the embedding during a *random walk* through embeddings produces with the trained model; starting from a vector that represents *cat*, measuring the distance until we reach the vector representing *dog*.

Let's think about retraining such a system, say, when a word meaning changes in the culture. For example, imagine that in an alternate reality, we live in a time where the word "cool" is only now starting to mean something other than "cold." New data instances added to the model would re-shape its topology, slowly bringing vectors that include the word "cool" closer to words like "trendy," "good," or "relaxed." I'm guessing that you might have had to take a moment to ask yourself whether cool does indeed mean trendy, good, or relaxed. Even writing the example, I am thinking something like, "I guess cool means those words, but it also means something else." I am thinking of the many senses of the word, imagining people who are known to be cool, thinking of cases where I use the word, asking myself whether things can be *truly cool*.

The platonic ideal of *cool* is something I can only really access through my own, internal random walk. It has no fixed instance, but exists between representations that I can conjure. When I say that "meaning is

[4] Although in some fields the word "data" is considered to be plural, in HCI the convention is to use as a singular collective noun, or, as in this case, an adjective that describes a type of object.

relationship," it is the cognitive process of both recalling and making new associations. The meaning is both in the previous state of the relationships between the word and instances of my experience, and the realization through explicit thought. It is also an emotional experience: part of my cognitive conjuration involves an aesthetic appraisal, gauging my cognitive experience against my in-the-moment feelings brought about by the memories and attempts at definition. This cognitive process is extremely difficult to bring to conscious attention: I don't actually known what I mean by "cool" until I have completed an appraisal process, but I can confidently use the word or even assess whether something is cool without having to think very hard at all.

We can use NLP model training and querying as a metaphor for our cognitive processes involved in meaning-making. What we currently know is like the current state of the model, expressed through associations. We can think of these very roughly like brain areas, neurons connected to other neurons. The model has weights between areas, which are like the density of neuronal connections between brain areas. Activating portions of the model is like activating parts of the brain; e.g., if I say to you "cold" and then "cool," the meaning for you will likely be in a temperature sense rather than the "trendy" sense. A major difference for our brains is that querying also reweights the model: for humans, running the software changes the hardware. That is, meaning isn't static but dynamic, even if some meanings are more *stable* than others.

However, the machine metaphor for the brain needs to be used with caution. We will talk about signal processing models of perception in the next section, but it is important to note a remaining meaning of *meaning is relationship*: personal relationships. Humans are social animals, and meaning as a personal experiential phenomena takes place in a social context as well as a physical environment that is perceived through meaning-making emotional processes which account for other people's experiences. Even as people become excited about artificial intelligence looking closer to what we want out of artificial general intelligence, machines do not have physical human bodies which have been designed over millennia to be in large social groups in the natural world. It is easy to forget in a cognition-focused academic context in a social media world where a lot of interactions happen on a computer, but humans experience a real, physical world at all times, no matter whether their attention is on a representation of a virtual environment. Even concepts such as logic are subject to the reality of our social brains. For example, the Wason selection task, a logic experiment, goes like this:

> You are shown a set of four cards placed on a table, each of which has a number on one side and a color on the other. The visible faces of the cards show 3, 8, blue and red. Which card(s) must you turn over in order to test that if a card shows an even number on one face, then its opposite face is blue?

Famously, most people cannot accurately choose which cards to flip over. However, rephrasing the task as a social problem radically improves the success rate:

> You are shown a set of four cards placed on a table, each card has an age on one side and a drink on the other. The visible faces of the cards show 16, 35, soda and beer. Which card(s) must be turned over to test the idea that if you are drinking alcohol, then you must be over 18?

Which did you find easier? The answer to the social framing is "16" and "beer" because you need to check that the teen isn't drinking and that whomever is drinking the beer isn't underage. It is identical to the even number problem, but I bet you would have to check carefully to confirm that, whereas you can become certain of the drinking age problem with less effort. The experiment illustrates many useful things about cognition, but for our purposes it serves as a live demonstration that our cognitive processes and experience of meaningful knowledge dependent on social framing, practiced external knowledge, and concrete examples. An experiential social metaphor makes the logic problem meaningful.

Lakoff and Johnson argue in their book *Metaphors We Live By* that cognition is essentially metaphorical (Lakoff and Johnson [2008]). What this means is that we make sense of abstract concepts by transforming problems into social, emotional, and/or physical frames that we have experience of and therefore can reason about concretely. For example, an abstract concept like the natural counting numbers are understood by relating the experience of matching our fingers to physical objects; any logical induction is ultimately rooted in this basic grounding in a physical, bodily reality. It's metaphors all the way down.

When it comes to building ML models of meaningful experiences, whether they are emotions or qualitative interpretations, knowing how to operationalize meaning is required. In many ways, in this dissertation, I take these wide-ranging theories of meaning and creates concrete demonstrations of how meaning can (or cannot) be measured and computed. After the work

at the hospice was put on hold, I tried to figure out how to study meaningful interactions with devices, but now without being physically present with people or robots. My attention turned to questions about how people mediated their relationships through electronic devices. I found a large repository of thousands of stories of people trying to navigate meaning through the messy relational dynamics of sharing their lives digitally, through online accounts, passwords, and digital devices. Quickly, the project became about how to meaningfully navigate this giant text archive while holding true to the foundational values of qualitative research.

We ended up building a system for exploring meaning in large document corpora, called Teleoscope. Taking the idea of expressing cognitive *schemas* through machine learning models, Teleoscope provides an interactive way for researchers to collaborate on a meaningful understanding of a large document set. Learning from my experiences trying to create meaningful interaction with robots, I focused on the improvisational, collective *process* of meaning-making, rather than attempting deterministic models of meaning.

In the end, this dissertation became my attempt to understand how the experiential reality of meaning can be understood from a design and computational perspective simultaneously with a theoretical and scientific understanding. Between an analysis of my projects with robots and detecting emotions and the design, evaluation, and analysis of Teleoscope, I hope to demonstrate an approach to meaning-making with computational devices that is both simple and theoretically rich. As such, I include a wide-ranging set of perspectives from philosophy, cognitive science, computer science, and human-computer/human-robot interaction.

1.2 Outline

The chapters in this dissertation will roughly follow a chronological structure. The dissertation was written as a series of four papers that I was first author on, two of which have been published, and two of which are under review. As stated in the Preface, this dissertation also includes reflections on other published work. For a roadmap of that work, see Figure 1.1.

First, Chapter 2 will discuss *meaning* as part of affective computing, and in particular, emotion recognition through signal processing. This involves the translation of biological signals into emotion labels. The signals represent biological measures such as heart rate, breathing rate, brain activity, and behaviours such as eye movement or touch gestures. Through this, we

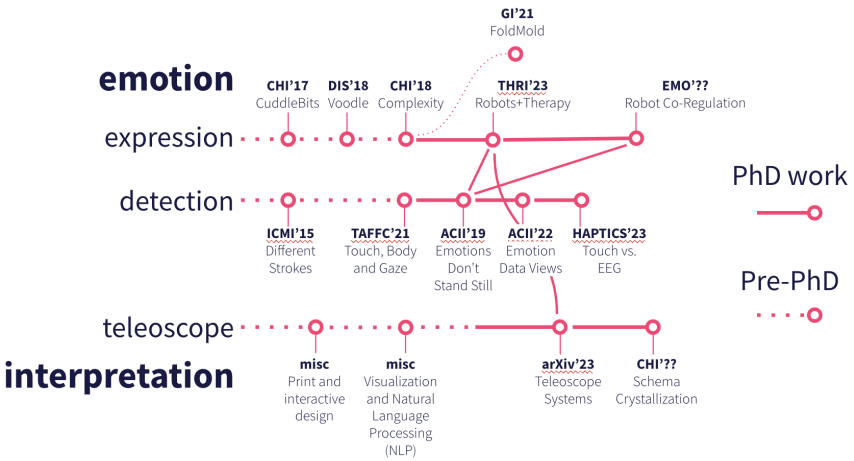


Figure 1.1: My dissertation comes from a wide range of work published before, during, and potentially after my PhD. In the above figure, each circle shows a "waypoint" as either a published academic work or significant project. Submissions that are in review have a "??" for the publication year.

will build up an intuition for low-level units of analysis in affective computing and how they can be understood as part of a larger system of emotion and meaning as a cognitive and bodily phenomenon.

Then, Chapter 3 will discuss *methodological* problems with the signal-processing approach, outlining how the experience of meaning may not be measurable, and suggesting alternatives to current approaches that respect the difficulty and complexity of meaning systems. This chapter and the previous will focus on examples from human-robot interaction, including electronics, actuation, and other physical device-oriented ways of thinking about meaning.

After discussing physical devices, Chapter 5 shifts focus to a web-based system that we designed for qualitative representations of meaning for large corpora in the range of thousands to millions of documents, called Teleoscope. Here, the problem of meaning is approached from a natural language processing (NLP) and thematic analysis perspective. The dissertation will discuss the design process which lead to Teleoscope, empirical studies with real qualitative researchers, and how engineering decisions dovetailed with Teleoscope's design goals.

Last, Chapter 6 will discuss a case study of using Teleoscope to collaborate on creating meaning structures with a real-world dataset. In it, we extend the value systems from the common qualitative research methodology of thematic analysis to the data curation stage, as necessitated by the scale of big data. We define "schema crystallization" as a way to describe the process of making sense of data both within one's own mind and using external representations, such as those produced withTeleoscope.

1.3 Thesis Statement and Contributions

In this dissertation, I report on, analyze, and draw conclusions from two multi-part projects that attempt to answer this question from different perspectives using interactive systems and machine learning. First, I look at computing meaning by attempting to detect emotions using signals derived from the body such as heart rate, brain waves, and gestures. Then, I look at computing meaning by making connections between documents to support thematic exploration of large document corpora.

My contributions in this dissertation are:

- A critical theoretical and methodological proposition for computationally representing, sensing and displaying real-time emotions.
- A synthesis of the theoretical and pragmatic basis of therapeutic care methods and their meaning for affective robotics, with an accompanying account of the constructed nature of emotions for HRI applications.
- The design and evaluation of a system (called Teleoscope) for capturing underlying meaning in documents through interaction with machine learning systems.
- An extension to thematic analysis for data curation to create meaning in large text datasets which we call thematic exploration, and a methodological concept of schema crystallization.

Through these projects, an underlying understanding of meaning-making as an embedded, embodied, emergent, interactive phenomenon is articulated. That is to say, meaning is *embedded* in a culture and environment, *embodied* in the whole of a person, and *emerges* through the process of interaction between a person, themselves, other people, and their environment. By understanding these epiphenomenal interactions, designers may be enabled to create computational systems that facilitate richer meaning-making.

1.4 Scope

Content from three published papers will be discussed where I am not the first author, but contributed significantly to the project they represent. The four papers written during my masters will also be discussed, but are not claimed as major contributions of the dissertation.

Papers that I was the first author on which were published during my PhD:

- Bucci, P. H., Cang, X. L., Mah, H., Rodgers, L., & MacLean, K. E. (2019, September). Real emotions don't stand still: Toward ecologically viable representation of affective interaction. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-7). IEEE.
- Bucci, P., Marino, D., & Beschastnikh, I. (2023). Affective robots need therapy. ACM Transactions on Human-Robot Interaction, 12(2), 1-22.

Papers under review that I first authored during my PhD:

- Bucci, P., Foord-Kelcey, L., Lee, P. Y. K., Singh, A., & Beschastnikh, I. (2024). Teleoscope: Exploring Themes in Large Document Sets By Example. arXiv preprint arXiv:2402.06124.
- Bucci, P., & Beschastnikh, I. (2024). Crystallizing Schemas through Thematic Externalization with Large Corpora. *Under review.*

Papers that I was involved in, have drawn on for a meta-analysis, but do not include as contributions for this PhD:

- Cang, X. L., Guerra, R. R., Guta, B., Bucci, P., Rodgers, L., Mah, H., ... & MacLean, K. E. (2023). FEELing (key) Pressed: Implicit Touch Pressure Bests Brain Activity in Modelling Emotion Dynamics in the Space Between Stressed and Relaxed. IEEE Transactions on Haptics.
- Cang, X. L., Guerra, R. R., Bucci, P., Guta, B., MacLean, K., Rodgers, L., ... & Agrawal, A. (2022, October). Choose or fuse: Enriching data views with multi-label emotion dynamics. In 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-8). IEEE.
- Cang, X. L., Bucci, P., Rantala, J., & MacLean, K. E. (2021). Discerning affect from touch and gaze during interaction with a robot pet. IEEE Transactions on Affective Computing, 14(2), 1598-1612.

As well as the following papers which were reported on for my Master's Degree and therefore are not part of the contributions in this dissertation, but do comprise part of my meta-analysis:

- Bucci, P., Zhang, L., Cang, X. L., & MacLean, K. E. (2018, April). 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 (pp. 1-11).

- Marino, D., Bucci, P., Schneider, O. S., & MacLean, K. E. (2017, June). Voodle: Vocal doodling to sketch affective robot motion. In Proceedings of the 2017 Conference on Designing Interactive Systems (pp. 753-765).

- Bucci, P., Cang, X. L., Valair, A., Marino, D., Tseng, L., Jung, M., ... & MacLean, K. E. (2017, May). 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 (pp. 3681-3692).

- Cang, X. L., Bucci, P., Strang, A., Allen, J., MacLean, K., & Liu, H. S. (2015, November). Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 147-154).

Chapter 2

Background: Measuring Emotion

In Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI) research, it is common to construct study paradigms as experiments with psychometric dependent variables. That is, using scales drawn from psychology that are designed to make psychological phenomena measurable, HCI researchers will often choose methodologies that position their interfaces as factors in an experiment, and then try to measure the impact of their interfaces on how people feel. Sometimes this can take the form of Likert scale questionnaires, where opinions are expressed in 1–5 scales from "Strong Dislike" to "Strong Like." In the case of affective computing (a subfield of artificial intelligence and often the domain of HCI research), these scales are often framed in terms of fundamental, universal emotions. A broad characterization of paradigmatic conflict in affective computing would be between *cognitivist* and *interactionist* approaches to understanding emotions.

2.1 The Rise of Cognitivism

Cognitivism itself is considered to be a reaction against behaviourism (Deigh [1994]), which is to say that early-century psychologists considered it only possible to build models of human affect out of observable behaviour and bodily feelings. Cognitivism rose out of the recognition that largely unobservable states of the mind also impact affect and states of the mind. That is, *cognitions impact cognitions* as well as produce feelings that we literally feel (rather than just think about).

B.F. Skinner is considered one of the great behaviourists whose methodological insistence was to produce experiments where direct stimuli association is the fundamental causal mechanic of interest. This mechanic, called operant conditioning (e.g., train a pigeon to press a button by giving it rewards of food), explains some phenomena quite well, but a cognitivist approach recognizes that a direct stimulus-response association does not

account for many human and animal cognitive phenomena and behaviours that we observe.

Cognitivism also begins to approach the mind with computation and information as part of theoretical models that were produced. Cognitive scientist Noam Chomsky famously did not like the dominance of B.F. Skinner's behaviourist approaches. In what is now considered to be the characteristic approach of the endearingly-named (or derisively-named) "Good, Old-Fashioned Artificial Intelligence" (GOFAI), Chomsky created formal mathematical models of language. These models were called generative grammars, and are still taught and used in Natural Language Processing (NLP) to analyze and generate language.

In contrast to behaviourism's simplistic stimulus-response model, cognitivism suggests that unseen computations are happening in a complex neural architecture that operates somewhat like the electrical connections in a computer. This is not to say that behaviourism is "wrong", despite whatever polemics of the day were written against behaviourism. But if one is interested in the underlying architecture of cognition, stimulus-response models are not complete, because they cannot account for the radically complex self-referential interactions that are happening within the cognitive (computational) architecture of the brain. The scientific impetus is therefore to construct formal models of cognitive processes that can be computed. A *symbolic* computational approach is taken where real-world phenomena are translated into symbolic independent variables that can be manipulated algebraically to predict dependent variables.

Due to the overwhelming dominance of a symbolic computational approach for modern science (and social science), it can be easy to miss why this was a radical shift in the scientific conceptualization of the mind. If behaviourists like Skinner were reacting against wild Freudian claims about the unobservable unconscious (or even more unobservable ideas like Jung's dream interpretation or the assertion of a world-wide intelligence dubbed the collective unconcious), the scientific paradigm now claimed that we could, in theory, model the unconscious observably on a computer. No longer are our minds a mysterious black box, but perhaps we can *see* the ghost in the shell. Further, the implication is that the logical relationship between real-world phenomena, mathematical models and computational simulations are close enough that a computer may, indeed, "think like we do."

Not all things are computable. Physicists and computer scientists developed computational theory to try to formally characterize the limits of computability. Computability theory tells us the time and space limitations of computation, and posits entire classes of problems that are not practical

to compute (at least, with our current paradigm of computing, not with a program that could finish before the potential heat death of the universe). Although physical models could be described with mathematical symbolic variables that account for infinite fields that span the universe, a computational model has to produce numerical solutions with finite resources and finite precision. The simulation of a model is existentially different than the algebraic formulation of the model: it has to be run on a machine.

With the relatively recent advent of machine learning, the cognitivist project has shifted from algebraic models to statistical models. Large language models (LLMs) have demonstrated that whether or not computers can "think like us," to some degree they can "write like us." Even before LLMs, cognitive scientists were debating whether statistical models were "doing science" in the sense of creating causal, predictive models of the universe. There is no doubt that they are making correlational, descriptive models of the universe: but is that science? Add to that our inability to explain machine learning models, it seems like there is a ontological-epistemic gap in contemporary machine learning approaches to science (we can say what is, but not really how we can come to know it). Given enough data, we can produce predictive statistical models of seemingly any phenomena that can be translated into bits and bytes, but it is actually quite hard to say whether those models are producing real knowledge about the phenomena they purport to describe.

The question as to whether cognitive processes are possible to faithfully model with computers is discussed extensively in the following chapters. The cognitivist approach to human-robot and human-computer interaction (HRI and HCI) remains dominant in the field. The fundamental assumption is that building explicit computational models of emotions is necessary for a well-functioning interface. The interactionist approach, as discussed next, stands in contrast to this.

2.2 Interactionism

Interactionism is a post-modern approach to understanding emotions within affective computing. The academic provenance of interactionism is from the sociological field of symbolic interactionism, which studies day-to-day interactions between people, places, and things. The interactionist research frame focuses on literal moment-to-moment interactions as the components of larger meaning-making systems that are reified in social institutions such as churches, schools, etc. Battarbee and Koskinen articulate this as a matter

of pragmatics in the socio-linguistic sense of the term:

> [The pragmatist] model is theoretical in nature, and shows that experiences are momentary constructions that grow from the interaction between people and their environment. In their terminology, experience fluctuates between the states of cognition, subconsciousness and storytelling, depending on our actions and encounters in the world. Experience is something that happens all the time: subconscious experiences are fluent, automatic and fully learned; cognitive experiences require effort, focus and concentration. Some of these experiences form meaningful chunks and become demarcated as 'an experience'—something meaningful that has a beginning and an end. Through stories, they may be elaborated into 'meta-experiences' that are names for collections of individual experiences. Even more recently, Wright et al. (2003) focused on what is common to all experience, describing four strands—the compositional, sensory, emotional and spatio-temporal strands—which together form experience. They also describe sense-making processes such as anticipating, interpreting and recounting. (Battarbee and Koskinen [2005])

An interactionist lens on affect would say that the emotional experience is not "located" in the individual so to speak. Instead, the individual is part of a larger system of interactions, and the emotional experience is happening in the interactions between them, other people, and their environments. Certainly people "feel emotions" themselves, but almost any part of the emotional experience is dependent on these micro-interactions with larger-scale systems.

As such, the pragmatic (in both the colloquial and academic sense of the term) approach in HRI/HCI is to focus on designing systems that support meaningful interactions. Methodologies such as *co-design* are favoured wherein researchers work with the exact target population of a system to incorporate their experiences of the technology directly into the design process as the product is developed. This is in contrast to a standard engineering approach of gathering requirements from a sample population, then assuming that the needs of the target population are close to the sample's needs. With co-design, the moment-to-moment interactions and relationships with the target population is considered vitally important, and as a result, the methods often include focus groups, workshops, and other community-oriented activities.

It may be easy to see why this approach is being advocated, particularly in the disability community. Technological solutions often need to be adapted so specifically to an individual that findings from a sample population simply might not apply. For example, there may be a demographic similarity between prosthetic users in the sense that they could say on a survey that they do indeed use prosthetics, but making a specific prosthetic comfortable and functional for any one person requires extensive customization. Further, an interactionist approach would recognize the meaning-making that happens through interactions with their physiotherapists, specific mobility barriers such as doors, or social activities such as sports where they use their prosthetics. Whether or not a prosthetic is "good" is a constantly-shifting idea that depends on both a design level and a society level. The designed form factor needs to be comfortable, but also the social level such as the building safety committees need to install doors that will open with the particular prosthetic they are wearing.

The observation for affective computing is that the phenomena we collect into the broad category of emotion may be similarly both individualized and based on moment-to-moment interactions with larger meaning systems. Deigh recognizes that emotions seem to have philosophical intentionality, that is, "aboutness."

> Intentionality is a property of actions and mental states. It is the property of being directed at or toward something. Emotions typically have this property. When one is angry or afraid, for example, one is angry at someone or something, afraid of someone or something. This someone, this something is the emotion's intentional object, that at or toward which it is directed. By contrast, bodily sensations of pleasure and pain, the comforting feeling of a warm bath, say, or the aching feeling of sore muscles, are not directed at or toward anyone or anything. They are not intentional states. Hence, a conception of emotion that identifies the phenomenon with feelings like these misrepresents it.

The interactionist approach importantly focuses on the intentionality of emotion by situating them in social contexts. For design projects such as the ones that are common in HRI/HCI, this is an incredibly important and oft-forgotten aspect of any system we create. As much as the interactionist approach helpfully focuses on meaning, it also does not account for the embodied and signal processing aspects of emotion that are both scientifically relevant and of particular interest to HRI/HCI researchers who care to cre-

ate physiological and behavioural sensing systems. Luckily, new work in the theory of constructed emotion helps to complete this picture.

2.2.1 Comparing the Theory of Constructed Emotion

As symbolized in a public debate at one of the top emotion computing conferences, the cognitivist point of view, typified by Rosalind Picard, is being challenged by the theory of constructed emotion, typified by Lisa Feldman Barrett. The theory of constructed emotion (TCE) recognizes that emotions are constructed out of a wide range of phenomena that are physically modelled in the brain.

> This process of managing the brain and body's energy needs, called allostasis, is based on the premise that a brain anticipates bodily needs and attempts to meet those needs before they arise...A critical feature enabling a brain to operate predictively is its ability to generalize or create higher level summaries from particulars. That means that the organism must constantly query its existing model about the current sensory array and what it is most like from its prior experience. Generalizations, also called abstractions, are constructed as an organism updates its internal model based on a wide range of highly variable instances over time. Predictions...are constructed from the organism's past experience with similar internal and external contexts. However, these contexts are never exactly the same twice. Predictions underlying particular instances of cognitions, emotions, perceptions, and actions will, therefore, be highly context-bound. (Fridman et al. [2019])

If the cognitivist metaphor for emotion is that the brain is like a computer running a GOFAI constraint network that produces a label probability for a given experience, TCE's metaphor is more like layers and layers of deep neural networks that manage a very large system of utility functions. Information processing is still happening, but the goal is to maintain allostasis across the entirety of the different parts of the body (which includes the mind). This challenges the cognitivist approach of labelling emotion:

> Emotions traditionally were, and in some quarters still are, commonly assumed to have an essence, meaning that all instances of a given emotion are presumed to have a core similarity either at a neural or physiological level (Barrett, 2006, 2017a,b).

> However, empirical evidence reveals a striking lack of consistency in emotional experience and expression, such that specific emotional instances (e.g., an experience categorized as anger or fear in a specific time and place) are highly variable across contexts, even within a person. Empirically, there are no biological "fingerprints" for specific emotion categories in the brain (Lindquist et al., 2012; Clark-Polner et al., 2017), in the face (Barrett et al., 2019), in the body (Siegel et al., 2018a), or in experience (Lindquist et al., 2013). That is, an emotion does not have unique and consistent physical features across individuals, or even within the same individual across instances. Thus, quite distinct experiences can be categorized (or labeled) as belonging to the same emotion category but still vary considerably in their features (e.g., whether anger is associated with a heart rate increase or decrease, or whether it is associated with a scowl on the face or not). Likewise, very similar affective experiences can be categorized and labeled as belonging to different emotion categories across instances (Lindquist et al., 2013). For example, it can be difficult to tell if you are feeling fear or excitement in advance of an important event (e.g., a marriage, the birth of a child, a major sporting event of your favorite team). (Fridman et al. [2019])

The implication is that making computer models that focus on labelling as a paradigm are at odds with empirical research on the alignment of emotion labels with actual biophysical signals. However, this is the current dominant approach in HRI/HCI, including in much of the work presented in this dissertation.

Part of the difficulty is that computer systems need to operate on clearly structured and labelled data. Similarly, for basic usability, users need clearly labelled concepts and interaction metaphors, and do not care whether or not they align with the biophysical reality of the underlying phenomena that the computer is attempting to model. Similarly, when attempting to communicate with a participant about an emotional phenomenon, we must use word labels such as "angry" because, while reductive, that is the commonly understood way of communicating about this complex phenomenon. How then do we reconcile the practical realities of communicating with each other and our computer models? The rest of this chapter focuses on problems of measurement both from a signal processing perspective, and from a communication perspective.

2.3 Breaking Emotions into Measurable Parts

Measuring emotions is a place where a lot of seemingly abstract philosophical questions can become quite concrete. One might have some skepticism over the utility of a first-year philosophy question: is *your* red the same as *my* red? Briefly, it is impossible to establish whether the subjective experience of the colour red is identical between people, since it seems that perceptions are entirely "private." You can't know what it's like for me to see red, just that we point to the same colour when asked which colour is "red." However, for emotions, the word label often corresponds to subjective experiences with vastly different descriptions. My "happy" is almost certainly not the same as your "happy."

Emotions can be thought of as a product of sensory perceptions, fast-acting reactions, slow-acting cognitive appraisals, and bodily reactions. The reactions can be voluntary or involuntary, attentive or inattentive. Emotion theories often contradict as to what actually constitutes or causes emotions. For example, some theories are body-first, stating that our experience of emotions emanates from a post-hoc appraisal of bodily reactions. Other theories are cognitions-first, stating that our thoughts produce our emotions. On a personal level, my experience of anger may be that I feel a flushed face, whereas others may feel their hair raising or even cold. Lauri Nummenmaa has a research programme wherein he asks people to colour in parts of their body which corresponded to a subjective emotional or aesthetic experience (Nummenmaa and Hari [2023], Blain et al. [2023], Ojala et al. [2022], Volynets et al. [2020], Torregrossa et al. [2019], Sushchenko et al. [2017]). Although there are commonalities, even across culture, there was by no means total agreement.

From a realist, objectivist perspective, if we were interested in measuring emotions, we would be interested in looking for common characteristics and physical components of emotions that could be verified by anybody with the same measurement devices. For example, Paul Ekman is famous for establishing that many emotional facial expressions are common across cultures (Ekman [1992], Ekman and Friesen [1971]). He developed a detailed system for categorizing facial expressions which gave rise to the now-common concept of micro-expressions (Ekman and Friesen [1978]). His claim is that there is a one-to-one correspondence between a micro-expression and an emotional state. He does not have to look too deeply at meaning or the subjective experience, because his definition of an emotional state is in terms of the objectively measureable signal, the facial expression.

In Ekman's system, any subjectivity in establishing the difference be-

tween emotions is essentially due to measurement error. There is a library of cross-cultural examples which have been discretely categorized into hierarchical clades of emotions by physical similarity. After extensive training, emotion assessors will reference the library to evaluate an image of an emotion. Any discrepancy between emotion assessments are equivalent to a statistical problem of determining inter-assessor reliability. The most frequent assessment is ground truth.

Many computer scientists have identified this as clearly fruitful ground for developing an emotion classifier based on images of faces. Computer scientists will use this large library of labeled data to train a classifier and test on, e.g., frames from popular films. The problems are: (1) do we trust the database? and (2) do we trust the theory?

Assuming that the data was faithfully collected, the theory falls apart if people can have categorically different emotions while having the same facial expressions. From an external assessment perspective, this happens to be the case. Many facial expressions are ambiguous if one is simply looking at another person, but particularly if the assessment is over a photo or a still frame of video.

There is also a problem with time extent. A micro-expression takes just a few milliseconds. If an emotional state is detectable in a single frame of video, then it implies that an emotional experience is decomposable into very small units of time. This doesn't seem to be the case: emotions take place over different time periods, with different expressive components, and, importantly for us, with different meanings (Ekkekakis [2013]). For example, if one feels happy and then sad, that is *meaningfully* a different experience than feeling sad and then happy.

There is great scientific value in an objective signal-processing approach like Ekman's. However, most emotion researchers want to know about emotions for a *purpose*, not just because the signals themselves are interesting (even if they are quite interesting). There needs to be a match between the claims of the science and the reality of the measurement. Ekman's work tells us a lot about facial expressions, and very little about subjective experiences. We want to know how somebody feels *inside* causes facial expressions. We would like to know that the word-labels that Ekman attaches to his images correspond to true universal experiences.

Such a causal assumption relies on a stimulus-response paradigm that does not hold for modern conceptions of cognition. For example, I cannot consistently evoke happiness by offering any person a coffee. The stimulus is too contingent: does the person like coffee? Did they already have a coffee? Did I offer them a coffee already? The level of analysis is inap-

propriate for an objective measure. I could look at the stimulus at a lower level by measuring heat or caffeine, but looking at the level of meaning and emotion would require a subjective measure, even if stimulus, response, and intervening machinery are objectively measurable. The reasons why they are not objectively measurable are interesting, and will be dealt with more fully with in a later section. For now, it is enough to say that the measurement devices are imprecise, and the intervening neural circuitry is formally chaotic and complex; that is, extremely sensitive to initial conditions and difficult to analyze in a modular fashion.

2.3.1 Constructing subjective measures

We are quite used to subjective measures through surveys, personality tests, and other psychological diagnostic tools. It is easy to lose sight of the reality that they have been constructed. Thinking further about our coffee example, while we *could not* establish a predictive, objective, causal relationship between coffee and happiness, we indeed *could* establish a descriptive, subjective, correlational relationship.

Some of the ways that psychological claims are misunderstood can be due to the failings of English. "Coffee makes people happy" is a cultural truism, but clearly false if taken literally (however unbelievable, it is the case that some people don't like coffee). Making the statement more precise, "coffee makes most people happy" is easier to defend, but begs contingency. "Of English-speaking coffee-loving adults that we studied, when served a fresh, hot cup of coffee in the morning, they were more likely to express pleasure-oriented phrases and facial expressions than not" approaches a measurable claim. Each portion of the phrase will be operationalized within the study design and execution. For example, "pleasure-oriented phrases" will likely be established through an open coding process, where study recordings are analyzed by researchers who transcribe the phrases and rate them according to their emotionality. During research team meetings, scales may be determined as to "amount of pleasure" so that researchers can decide which instances meet a determined threshold of pleasurability.

When we say that subjective measures are "constructed," the coffee example above gives us an idea of how the construction process happens. There are objective portions of the coffee pleasureability scale, such as spoken phrases, which are established through a process of subjective interpretation by researchers while reviewing the data. It is easy to miss that even this part of the study is culturally-contingent: if a group of non-English speakers reviewed the same footage, they might come up with different numbers.

Further, the determination of the scale during the research meeting is part of the construction: each individual researcher might review the same footage and get different numbers without it. A frame must be imposed to ensure consistency.

This is why the claim is correlational and descriptive, not causal and predictive. The framing of the study really does make it sound like coffee causes happiness, but it is only within the extremely narrow control of the study premise. Change almost anything about the premise to be outside of the cultural expectations, and you might find wildly different results. Put the participants in a sauna, take off their pants, change the time of day to 2am, and the results may be very different. In fact, move out of a coffee-drinking country to a tea-drinking country and the results may be very different. The coffee does not cause the happiness, the entirety of the cultural situation creates the happiness and the coffee is nothing but a concrete external marker.

Just because something is correlational, does not mean that we cannot get anything useful done with it. The coffee study would be useful to a market research company. Psychology does a similar thing with mental health diagnoses with considerable effectiveness. For example, psychological diagnoses such as Borderline Personality Disorder (BPD) are important and useful psychological constructions. BPD was constructed over a decades-long process where clinicians noticed that certain patients had behavioural expressions that seemed like psychosis, but were still grounded in reality. True psychosis can involve true hallucinations, however, patients with BPD instead seemed to be offering extreme interpretations of real events that bordered on psychosis (hence the name *borderline* personality disorder). Over the subsequent decades, as more case reports came in and were reflected on, the diagnostic criteria were added to help differentiate from true psychosis, and BPD was added to the Diagnostic and Statistical Statistical Manual of Mental Disorders (DSM) in 1980.

Having a diagnostic label such as BPD helps to direct treatment. For example, anti-psychosis drugs may not work for patients with BPD; similarly, the gold-standard treatment for BPD, dialetical behavioural therapy (DBT), may not work for patients with true psychosis. BPD seems to describe a truly-existing phenomenon. But is BPD "real" or "constructed?"

We would still say that BPD is "constructed." For example, it is difficult to differentiate BPD from certain kinds of post-traumatic stress disorder (PTSD), particularly because there seems to be an extremely high comorbidity. There may be an underlying phenomenon which gets expressed differently depending on cultural context. Or the diagnosis may describe a

collection of potentially different diagnoses that will become clear through decades more of observation. A really-existing underlying phenomenon may even be a fiction—we simply do not know.

For the moment, scales have been constructed to measure aspects of BPD, such as the Difficulties in Emotion Dysregulation Scale (DERS). Scales such as the DERS aid in making diagnostic determinations, but they are not definitive, and are highly subject to individual variance. Items on the scale are often constructed as Likert Scales with 1–5 answers:

> *Item 8: When I am upset, I feel out of control*
> 1=Almost Never
> 2=Sometimes
> 3=About Half the Time
> 4=Most of the Time
> 5=Almost Always

Multiple items are combined to create scores for different scale factors. Scoring high on the DERS does not alone guarantee that a patient has BPD, or even that they have difficulties in emotion deregulation. The scale must be used in conjunction with observation, interpretation, and treatment to determine a "true" diagnosis, which may never happen.

Although the DERS is widely cited at over 11,000 citations, the way in which it was constructed was somewhat ad-hoc. From the original DERS paper:

> The DERS items were developed and selected on the basis of numerous conversations with colleagues well versed in the emotion regulation literature. The NMR (Catanzaro & Mearns, 1990; see later) was used as a template and helped to structure the format of some of the items (although not the content of the items). Specifically, in order to assess difficulties regulating emotions during times of distress (when regulation strategies are most needed), many items begin with "When I'm upset," similar to the NMR.

The DERS was then academically "validated" by administering to just over 600 undergraduate students in two studies. In practice, the scale has been validated through clinical adoption, further scholarly investigation, and adaptation into new cultures and languages. It is more a very useful cultural artifact than it is an objective measure. Certainly, the DERS certainly

describes something important about the human experience and uses statistical validation to establish efficacy. But the purpose of the scale is to aid diagnosis and treatment, not to make an ontological claim. You cannot be said to "have" emotion dysregulation if you score high on the scale, you can simply be a likely candidate for diagnosis and treatment.

The reason for discussing subjective construction at length is that the sense in which subjective measures are constructed are quite different than measures in signal processing and computation. One cannot be said to "know" that things are "real" in the same way as to say that a chair is "real" or a signal is "real." It may sound like philosophical nit-picking, but, for example, my students often pitch projects to me where they plan to make something like a depression detector which monitors Zoom for trends in facial and behavioural markers. Explaining that it is practically undesirable to make such a machine and perhaps categorically impossible is difficult. Many HCI researchers make this exact same mistake. This is due to the misunderstanding of what is possible to detect with a signal vs. through an interactive, interpretive process. In the next section, we will explore what we can detect from a signal.

2.4 Emotional Signal Processing

In Computer Science and Electrical Engineering, a signal refers to a variety of physical phenomena that can be translated into an electromagentic wave using an electronic device. For example, sound waves propagate through the air as pressure which flex the diaphragm of a microphone, which in turn moves a magnet through a coil of copper wire, generating an electric current in the wire. Physical motion is translated into electrical signal. For digital devices, the signal is encoded as binary bits, usually with some loss of precision since a continuous analog signal will need to be sampled and discretized into a finite number of bits.

To generalize the phenomenon, computer scientists (and physicists) simply refer to signals as information, which can be represented in bits. All physical phenomena are therefore reducible to bit representations since it is assumed that all physical phenomena can be sensed in some manner. In fact, the contrapositive is that the only phenomena which we can consider to be *physical* are that which *can* be sensed. Both colloquially and technically, signal is therefore anything that can be sensed, and anything that can be sensed is a signal.

Although this may sound reductive, it is simply the way that compu-

tational systems must work. There is no way to make a computer operate on any information except to input it as an electric signal. This is what is meant by an emotional signal: any physical phenomena having to do with emotions that can be translated into digital electrical signals. The fundamental assumption of affective computing therefore is that emotional signals exist and can be processed using electrical systems and computer programs.

For the purposes of this dissertation, we will categorize emotional signals as follows: (1) Behavioural signals, i.e., things we can choose to do, which are generated from voluntary actions such as skeletal muscle movements; (2) Somatic signals, i.e., things in our body that we don't choose to do, which are generated from involuntary actions in the body such as heart beats; (3) Activation signals, i.e., things in our nervous system, which are generated from neuron activation such as brain waves.

These categories are offered as conceptual aids, not as precise delineations. For example, we think of breathing as mostly involuntary, but we can also control our breathing when we choose to. Similarly, eye movements are full of essentially involuntary saccades, even if we direct looking voluntarily. To further differentiate these cases, we will say that behavioural signals might also be attentive or inattentive.

Behavioural signals can be captured through a wide variety of sensing media and signal processing techniques. Body motion can be captured through video, with the most advanced techniques combining visible light through normal cameras with infrared light which capture depth. Or, breathing can be measured through a stretch sensor wrapped around the abdomen, where the piezoelectric effect translates deformation into electrical signal.

Somatic signals often need to be sensed through an indirect measure. Heart rate can be sensed through electrical signals propagating through muscle tissue, but can also be inferred through blood oxygen contents, which itself is measured by analyzing the absorption spectrum of light shone through blood via the skin.

Activation signals are also indirect. Individual neuron activation is mostly not possible to sense except when tissue has been removed. Instead, we have large volume measures such as Electroencephalography (EEG) which measures brain waves through electromagnetic perturbations at the scalp, or, the current gold standard of functional magnetic resonance imaging (fMRI), which measures blood flow in the brain that is correlated with neuron activation. The spatial resolution of both of these technologies is surprisingly low: EEG is often in the range of 2–128 electrode locations on the scalp and can only really capture the top cortical layer of brain activation, and fMRI

is just under $4mm^3$ (roughly fifty to one hundred thousand neurons).

When we talk about each of these types of emotional signals, there are certain assumptions that we must work with. For example, vagus nerve activity is associated with emotional state, but very difficult to directly measure. A somatic signal such as galvanic skin response (think: "sweatiness" measured by electrical conductivity of the skin) can be used as a proxy; so can heart rate variability (HRV), which is a measure of small changes in time between heart beats. Higher HRV means better vagus nerve tone which means a more relaxed subjective state. One can therefore imagine that a measurement of HRV may be a measure of relaxation. And while there is some truth to that, it is sort of like saying that measuring water velocity is a measure of boiling. Not entirely wrong if we accept that boiling must increase water velocity to some degree, but realistically, the measure is not categorically aligned, and certainly not decoupled from many other potential confounds. In this next section, we will talk about establishing the relationships between signals and emotional states.

2.5 Relating Subjective Measures with Emotional Signals

Externalising subjective experiences is difficult. At the extreme end of the philosophical spectrum, solipsism tells us that we can only possibly prove the existence of our own experience. At the other end, objectivism tells us that we can only possibly prove the existence of that which is externally observable. Bridging the gap comes in two broad forms: (1) a statistical approach (post-positivism) which we will discuss here; and (2) an interpretive approach (constructivism, post-modernism, etc.) which we will discuss later.

A post-positivist (statistical) approach may look like constructing scales (as from above) concerning emotional experiences. Items on a scale would indicate the category and intensity rating of emotional experience. The frequency with which a particular body signal is associated with a category and intensity rating would indicate a strong correlation between signal and rating. For example, in Ekman's work, if people consistently chose a particular facial expression to be labeled with the word "happy" at a rating of "3/5", we might confidently label the facial expression as "moderately happy."

It is worth noting the statistical assumptions: most people will interpret the scale items similarly enough to produce a unimodal normal distribu-

tion centred on the true population value, i.e., by the central limit theorem (CLT), the average rating closely approximates what we would expect if we could run the experiment on every living person. E.g., almost everyone in the study rated the "happy" facial expression near to 3/5, and when we use the scale again, everyone else will also be in that range.

Restating the previous to bring out a nuance of the corollary: error is accounted for in the experimental design such that it is evenly distributed around the mean and is either (a) accounted for in the blocking design or (b) has no hidden variable as expressed by a random distribution. E.g., men and women don't read the question differently, and if they do, the statistical significance of the item still holds when we split the experimental analysis by that demographic block.

Putting the previous two paragraphs more simply: every statistical approach is dependent on a large group of people rating the same thing the same way, and assuming that any new people who use the scale would belong to the same distribution of people as were in the study.

To support these approaches, categorical emotion theories have been developed along with corresponding scales and measurement instruments. Plutchik's emotion wheel posits a conical emotion space where core emotions are arranged like petals of a flower, and emotional intensity increases away from the point of the cone (which represents neutral emotion) (Plutchik [1982]). Eschewing the language of "emotion," Russell posited that below our surface-level emotions exists a 2D circumplex *affect* space of "arousal" and "valence" onto which all emotions can be mapped (Russell [1980]). Researchers have operationalized this theory into a variety of study instruments, including the Positive and Negative Affect Schedule (PANAS) which includes words that map to different areas of the circumplex, and the Self-Assessment Manikin (SAM) (Watson et al. [1988a], Bradley and Lang [1994]). The SAM may also include a 3rd dimension, called "dominance" or "power," which helps to differentiate low-valence, high-arousal emotions like "fear" (low power) from "anger" (high power).

Affective computing studies often have a design where the researchers choose a scale, often derivatives of Russell's theory, and attempt to relate emotional signals to scale items. Participants will use something like the SAM to choose how a stimulus makes them feel. Then, the stimulus is established to be correlated to a rating. For example, a scary movie clip may be shown to a participant, their heart rate may increase, they rate their experience of the movie as "low valence" and "high arousal," and the heart rate signal is established as correlating to the emotion rating.

The question for the reader is: did that experiment create an emotion

detector? Keep the question in mind as we first discuss interpretive approaches, and then the neurobiology of meaning.

2.6 Interpretive Approaches to Emotions

Although academic psychology is largely focused on experimental paradigms, clinical psychology is more grounded in interpretive approaches. This may not be surprising: despite having died almost 85 years ago, Freud's approach to understanding the unconscious is still deeply embedded in our cultural understanding of psychology. Even if the field of psychology has developed greatly, interpretive approaches and talk therapies as pioneered by Freud are the fundamental basis of therapeutic interventions today.

The interpretive answer to the question posed in the previous section would be a definitive "no." The assumption with interpretative approaches is that emotions have meaningful content, philosophical intentionality, affective richness. That is to say that emotions are produced not just through direct experiences of stimuli, but also a tapestry of meaning woven through narratives that include memories, somatic (bodily) experiences, direct environments and cultural situation. An emotion is "about" something as much as it is happening "because of" the place that you are in, the people around you, and what you've experienced before. Reducing it to a scale or a stimulus response is removing most of what people think about when they say that they feel an emotion.

Interpretive approaches can at first glance seem unscientific because they do not make claims about causality or controlled variables, and do not value generalizability. However, interpretive approaches have their own definitions of rigour in terms of research methodology and academic contribution. Some tend to be analytical, establishing rigour via the proper application of theory. Others tend to be empirical, establishing rigour via the proper application of methodology.

For example, a classic narrative approach would understand emotions as being closely linked to stories that we tell about ourselves. Words such as "happy" would not have much meaning without being substantiated by stories that articulated the sense of the word, e.g., happy in the context of a child winning a carnival game is different than happy in the context of drinking a warm cup of tea on a winter night.

A more theoretical, analytical approach would look at cultural metanarratives of happiness, typically with reference to social roles and identities. For example, one might understand happiness further by analyzing

what it means to be a "mother" who is happy, because being a mother is a cultural archetype that carries a set of culturally-determined meanings that an individual may be reacting to.

The way in which interpretive claims are made depends on the methodological tradition of the practitioners. Some qualitative analysis skews very closely to classic post-positivist study design where the data is coded with strict mechanistic guidelines, and statistical analysis on codes is where claims are drawn from. For example, "participants most often used 'happy' words" is a statistical claim. Or, using a non-statistical hermeneutical approach, interpretations must be theoretically supported, and frequency would not matter. Grounded theory would be yet another approach where "small theories" are developed by interpreting data participant-by-participant until theoretical saturation[5] is met (i.e., diminishing returns).

The question from the previous chapter about whether we made an emotion detector depends on whether one believes that the interpretation is *part* of the emotion. If so, then we should be accounting for interpretation even in signal-processing paradigms. It's not a matter of leaving interpretation to the artists and signals to the scientists; the science must account for the physical reality of meaning.

2.7 Neurobiology of Meaning

To think that meaning has physical extent may be surprising. However, a grounded look at the neurobiology of perception gives us clues as to how to think about meaning as a cognitive process. Therefore, this section will build the intuition for how to think about the brain as a meaning-making organ.

Many cognitive scientists account for cognition starting with perception. Visual perception is commonly studied; we will use examples from eye movement and visual illusions to establish the impact of meaning on eye mechanics. We will also draw examples from touch perception, itself an incredibly complicated phenomenon.

Let's trace a touch through the body. Kryklywy et al. performed a study where "social" touches were applied to participants, either a somewhat painful pressure on the thumbnail or a pleasant caress with a brush on the forearm, while fMRI data was being taken (Kryklywy et al. [2023]). A broad systemic trace would look like:

[5]Saturation is also referred to as information power to liken it to statistical power, getting at the same concept of diminishing returns on new data.

(1) Pressure would be sensed simultaneously by a variety of touch sensors (Merkel cell, Meissner corpuscle, Pacinian corpuscle, Ruffini endings, etc.) which initiate action potentials along a chain of neurons. (2) Action potentials are processed through the dorsal root ganglion, essentially a nerve junction in the spine. (3) The signal may be further processed by interneurons (e.g., to initiate a reflex response), but ultimately passed through the spine to the brain. (4) The thalamus would process the signal and relay it to the appropriate parts of the brain. (5) The sensory cortex would receive the signal, and the participant would consciously experience the touch.

Although this is a good general path description, Kryklywy et al. found that there was a differentiation between the painful and pleasant touches in terms of the speed and location of the processing. That is, different neural pathways seemed to be involved depending on the quality of the touch. They hypothesized that the pleasant social touch may be *faster* than the painful touch, potentially due to increased myelination along the pleasant touch pathways.

This is an illustration of how our bodies are mechanically constructed to process different meanings differently. An emotional signal may literally be processed differently depending on whether it is painful or pleasurable, not by some magic due to culture, but because of the neuronal wiring involved. However, our next two examples complicate this, as they illustrate how meaning can change perception.

First, take a moment to convince yourself that you can change your perception simply through mental effort. Figure 2.1 is a classic illusion depicting a wireframe cube (Kornmeier and Bach [2005]).

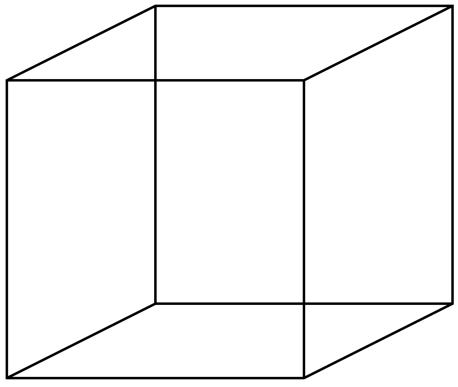


Figure 2.1: Is the cube sticking out of the page?

Is the left or the right face "sticking out of the page"? Take a moment to see whether you can convince yourself of one or the other. Then, see if you can switch it. Some people can eventually find a way to switch between perceptions at will. Nothing about the physical world outside of our bodies have changed when that happens. This is purely brain state impacting perception.

Context can also impact perception. A classic example is that colours are perceived not as absolute values, but instead relative to other perceived elements. In the following diagram, each colour of grey is identical in colour value, but differs in perception due to the perception of shadows.

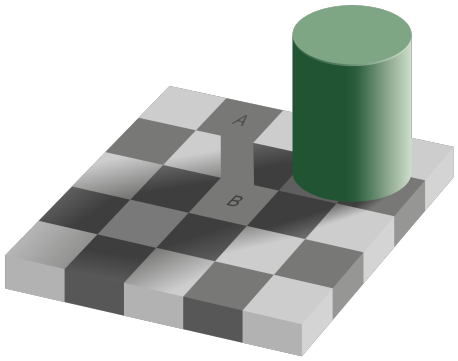


Figure 2.2: The colours of grey are identical, but seem like a "white" check inside of the shadow and a "black" check outside of the shadow.

The general phenomenon points towards the premise that we psychologically construct our reality as much as directly perceive it. Gestalt psychology demonstrates that wholes are perceptually made from parts. Our perceptual experience includes "filling in" values that are not directly causally linked to an immediate real-world phenomenon.

However, our perceptual systems are even further primed by meaning. Studies on visual attention show that meaning impacts the speed of looking. That is, if we are shown something of social value, like money, we will look at it faster than something without social value.

That should be a surprising result. Money is not something that physically evolved with human beings. We very likely do not have an inbuilt money detector in our visual systems in the same way that we likely have inbuilt snake detectors. It is entirely something that is socially constructed to have value. The fact that we look faster at objects with money than not tells us that our body movements are impacted by meaning. Meaning is not

an abstract thing, it concretely changes the way that our body moves.

Eye movement offers a particularly interesting microcosm of this. Our eyes are essentially "programmed" to look at everything in our visual field through an entirely involuntary neuronal signals. The experience of looking involves many physical eye movement deviations called saccades that we do not perceive. The act of voluntarily choosing to look at something is effectively a lower frequency suppression of the involuntary movement programming. The emergent effect is that we perceive ourselves as having a strong free-will choice to look, but it is better understood as a continually evolving process of involuntary looking being suppressed by voluntary looking. For an intuition of how this works, think about a cat trying to ignore a moving string. They can try to look away, but are clearly unable to fully voluntarily control their attention system. Looking at meaningful things will essentially be programmed both into the underlying involuntary looking as well as the voluntary looking.

2.8 Meaning Impacts Emotion Rating

In my final paper of my master's thesis (Bucci et al. [2018]), we ran a study where people were asked to rate the emotions of a robot behaviour: was the robot happy, sad, etc., given a particular breathing behaviour. The conclusion of the study was that people constructed narratives about the robot which far exceeded the behaviour of the robot. For example, people would, without prompting, pretend that multiple robots were present. Or, they might believe that the robot was lying or hiding something from them.

We had had a suspicion that this might be the case for some time, which is why we ran the study. When running a co-design study called *Voodle* where voice actors would puppet the robots using their voices, it was difficult until the participants decided on a narrative frame for the robot, e.g., "the robot is like my dog." To understand the task, we had to create a concrete example of something tied to the participant's real experience of life as an example. The meaning was not derived from an abstraction, but from an analogy.

Prior to this work, we had run as study where people told a stationary robot about emotional experiences while interacting with it via touch. We were surprised at how effective the robot was at facilitating people to tell their emotional stories. Combining our previous hypothesis that we could detect emotions through touch, but wanting a higher-fidelity sensor, we started to devise an experiment where we attempted to discern emotional

state through EEG signals.

A long study design process began. We started by telling each other emotional stories while petting the robot ourselves. Our thought was that we would be able to accurately label our own emotions by playing back a video recording and labeling with the SAM. However, we quickly discovered that the data was essentially unrepeatable.

Designing a study task with brain data requires extremely repeatable tasks that have precise timing. We searched for an appropriate emotional stimulus. Videos are often used to elicit emotions, however, we could not convince ourselves that we could guarantee that our participants were truly having comparable emotional experiences with the video clips. First, the clips were short, and we questioned whether someone could really be immersed enough in the emotional experience in a 30s or less clip. Second, despite the clips being validated, we did not find them particularly evocative ourselves. We asked: how can we be sure that someone is feeling something consistently?

Eventually, we decided on playing an evocative video game with repeatable sections. After play testing tens of video games, we decided on one that (a) had very simple keyboard controls; (b) very clear emotional events that could be timestamped; and (c) consistently elicited emotions for ourselves and our large research team while playing.

Learning from our previous robot studies, it seemed that we would need a unique way of measuring emotions for the EEG study. If we wanted to be able to reproduce an emotion label and intensity from an EEG signal, it would require millisecond-accurate timing, consistent meaning of labels and intensity ratings across participants, and multiple samples per emotion label and intensity. The task was enormous and our eventual study design was complex.

The theoretical underpinnings to the emotion labelling task are reported on here with an adaptation of our published paper "Real Emotions Don't Stand Still: Toward Ecologically Viable Representation of Affective Interaction." The results are reported on in two published papers that are not part of this dissertation, "FEELing (key) Pressed: Implicit Touch Pressure Bests Brain Activity in Modelling Emotion Dynamics in the Space Between Stressed and Relaxed", "Choose or Fuse: Enriching Data Views with Multilabel Emotion Dynamics," however I will summarize and discuss the results in the subsequent section.

Chapter 3

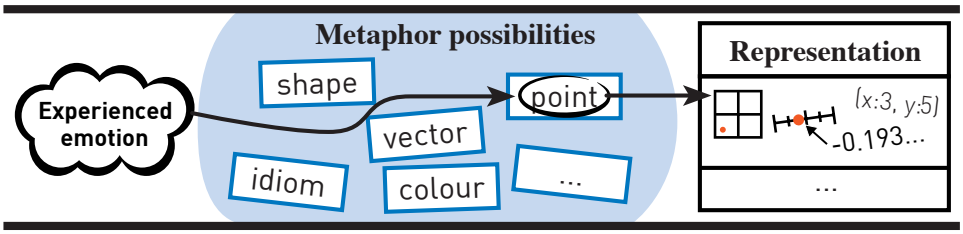
Real Emotions Don't Stand Still

This chapter presents the published paper *Real Emotions Don't Stand Still: Toward Ecologically Viable Representation of Affective Interaction*, published in 2019. It presents an early attempt to justify and theorize about our methodological approach in the EEG study, which we considered to be a significant departure from common affective computing emotion sensing studies. It argues against the idea that emotions can be fully captured in a state, and presents potential alternative formulations of Russell's affect grid which respond to attempts to use the affect grid as equivalent to a control space for robot motion. Further, it points out the difficulty with specifying the narrative perspective that can be confused in robot emotion evaluation tasks.

3.1 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 (Yohanan and MacLean [2011]) 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



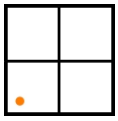
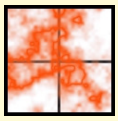
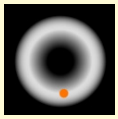
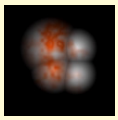
Metaphor description	One representation	
A point is singular, dimensionless, and infinitely precise. Representing emotion states as points may have unintended topological implications—even using a point to represent an average emotion measurement is problematic (Table 1).		Point (a)
Conceptualizing an emotion as an area opens up the possibility of considering multiple, coinciding emotions in our computational models. An area metaphor can capture conflicting, simultaneous emotions.		Area (b)
We must also consider the space over which an emotion may take place. A non-linear space with contour expresses the probability of emotion states and could express the predicted trajectory of emotions evolving over time.		Nonlinear (c)
The state space might also evolve over time. Shown is a space where, for a particular person in this moment, high-arousal, low-valence is most likely, but they might describe an ambivalent arousal.		Mix (d)

Figure 3.1: 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).

effort, perhaps even formal training, to both hold back our reactive instinct and articulate our emotions.

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

I. *Common metaphors do not account for dynamic emotions.* Representing emotions that change over time, are uncertain, or are in conflict requires amending our current metaphors and representations of emotion.

II. *Contemporary practices do not always explain whose emotion is being measured.* Interaction framing is often unspecified, leaving uncertainty in what an emotion is being ascribed to: a robot's behaviour, a participant's response to the behaviour, or something else.

III. *The meanings of measurement scales are ambiguous.* We often fail to create a shared understanding of measurement scales between participants and researchers.

3.2 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 (Lakoff and Johnson [2008]). 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 3.1).

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 Lakoff and Johnson [2008]).

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 3.1, top right) Russell [1980]. 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 (Barrett and Russell [2014b]). For example, to communicate with participants about their emotion, we can employ instruments such as the Affect Grid (a discretized 2D circumplex) Russell et al. [1989] or the Self-Assessment Manikin (SAM), which splits the arousal-valence-dominance space into three scales with cartoons for each scale item (Bradley and Lang [1994]).

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/generality. 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.

3.3 Related Work

Recent theoretical work in emotional interaction has challenged the dominant "signalling paradigm" (Jung [2017]) 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 (Jung [2017], Leahu and Sengers [2014]). 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 (Jung [2017], Bucci et al. [2018], Bakhtiyari and Husain [2014], Hollenstein [2013]). 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

Table 3.1: Dimensional theories of emotion use the metaphor of multi-dimensional scalar quantities to reason about subjective experiences. This table outlines the implicit assumptions and consequences of strictly interpreting emotions as a point on a linear, dimensional space regarding Assumption 1. This table elaborates on *Problem 1* from *Related Work*.

Implicit Assumption 1: Emotions can be represented as a single point-like state			
Implication of making assumption	*Ensuing representation limitation*	*Example of experience mismatch*	*Representation or experience mismatch*
Focus: 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.
Fixedness: 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.

Table 3.2: Continuation of dimensional theories of emotion discussion focusing on Assumption 2. This table further elaborates on the implications of strictly interpreting emotions within a continuous and linear space as discussed in *Related Work*.

Implicit Assumption 2: Emotion space is continuous and linear			
Implication of making assumption	*Ensuing representation limitation*	*Example of experience mismatch*	*Representation or experience mismatch*
Linearity: Emotions must be spatially distinct; linear, equidistant points correspond to similar sizes of emotion diffs.	Difficult to convey the size of qual. diffs 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 size of effort?	Default emotion rating scales are linear/uniform. But, not all perceptions are linear (e.g., perceptions of loudness are exponential).
Probability: 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/are dependent on prev. 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).
Unclear Temporality: If time is allowed, instant transitions between extreme emotion states are not representable.	Traversal from one emotional state to another can feel instant and 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?

converged upon as a result of continuous interaction (or "emotion coordination") (Jung [2017]). 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 (Fischer et al. [2019]).

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 (Bretan et al. [2015], Fischer et al. [2019], Ekman and Friesen [1971], Jung [2017]), Russell's dimensional model of affect (Song and Yamada [2017], Bucci et al. [2017], Saerbeck and Bartneck [2010], Nakata et al. [1998], Bhuwalka, Kunal; Icel, Nur; Gong [2018]) or a combination of both (Saldien et al. [2010], Yohanan and MacLean [2011]). Instruments include the Affect Grid (Russell et al. [1989]), the Self-Assessment Manikin (Marín-Morales et al. [2018], Saerbeck and Bartneck [2010]), or the PANAS scales (Bakhtiyari and Husain [2014]).

Herein lies the dilemma: computational models of behaviour require labels, but behaviours cannot be consistently and directly labeled with a single emotion (Leahu and Sengers [2014]). We could add contextual details to computational models to improve labelling accuracy (Breazeal et al. [2013], Damm et al. [2011], Bucci et al. [2018]). Alternatively, we could actively choose to represent conflicting or mixed emotions, aligning more closely with how behaviours are experienced and interpreted in real life (Bucci et al. [2017]). We present a discussion of alternative representations in Section 3.4.

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 (Hodges [2008]). 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 (Int [2018]). 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 (Shen et al. [2018], Strohkorb Sebo et al. [2018],

Williams et al. [2018]) as excellent framing examples. Robots are introduced as situated in the task, participants can conceptualize the interaction prior to rating, and experimenters listen to and iterate with participants to establish meaning.

Fortunately, there are ways to avoid this situation without evident compromise of scientific rigor. Some HRI studies implicitly explicate frame by asking contrasting questions using different frames (Bretan et al. [2015], Nakata et al. [1998], Bucci et al. [2017]). Others establish frame through clarifying interviews where participants explain their interpretation of the study task (Bucci et al. [2018], Leahu and Sengers [2014]). Still others use concepts from theatre. Bucci et al. establish roles, characters, and settings for an interactive scene (Bucci et al. [2018]). Westlund et al. do this through an interactive theatrical process (Westlund et al. [2017]): 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 (Marino et al. [2017]).

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 (Sullivan and Artino Jr [2013]). 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 (Breivik et al. [2008], Stinson [2009], Price et al. [2018]).

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

of qualitative and quantitative research designs ("mixed-methods") that include participants directly in the co-construction of meaning: collaboratively understanding the rating scales (Boehner et al. [2007], Jung [2017], Gao et al. [2017]). 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 3.6.

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.

3.4 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 (Tables 3.1 and 3.2) 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' Barrett and Russell [2014b]. 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.

3.4.1 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 3.1.

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 the-

oretical 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.

3.4.2 Nonlinear spaces: topography of 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 3.1 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.

3.4.3 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.

3.5 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 (Jung [2017]).

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 3.3, 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.

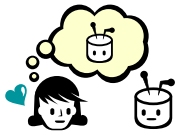
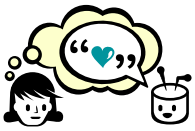
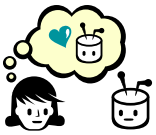
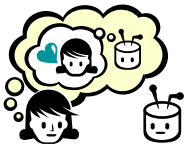
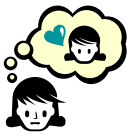
Many of the instruments we employ were originally designed for self-report of one's own affective state. For example, the SAM is intended as an easily understood, culturally universal method for a participant to express their internal affect via cartoon depictions of the body (Bradley and Lang [1994]). 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.

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.

Table 3.3: 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 *evidence* 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.

3.6 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.

3.7 Conclusion

In this chapter, 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 4

Moving from the EEG Project to a Deeper Understanding of Meaning

The previous chapter was our position paper which outlined the theory of our approach in the EEG Study. This chapter outlines the actual approach we took, summarizes the two resulting papers for which I am not a first author, and presents the final published critique paper for which I was a first author.

- Bucci, P., Marino, D., & Beschastnikh, I. (2023). Affective robots need therapy. ACM Transactions on Human-Robot Interaction, 12(2), 1-22.
- Cang, X. L., Guerra, R. R., Guta, B., Bucci, P., Rodgers, L., Mah, H., ... & MacLean, K. E. (2023). FEELing (key) Pressed: Implicit Touch Pressure Bests Brain Activity in Modelling Emotion Dynamics in the Space Between Stressed and Relaxed. IEEE Transactions on Haptics.
- Cang, X. L., Guerra, R. R., Bucci, P., Guta, B., MacLean, K., Rodgers, L., ... & Agrawal, A. (2022, October). Choose or fuse: Enriching data views with multi-label emotion dynamics. In 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-8). IEEE.

The game we chose to use in our study was an atmospheric side-scroller called "Inside," where players control a boy who is attempting to escape from people who hunt him down. The game is organized into scenes with save points that represent the beginning of a leg of the escape journey. The player can choose to run left, right, or jump. Lighting in the game is dark and the player is often occluded by objects in the foreground, creating plenty of opportunities for scary moments. The affective range is somewhere between creepy, to tense, to quite stressful. Moments of relief are potentially

possible when the character is able to hide, outpace, and successfully escape his pursuers. Unlike other games, fighting back is not possible.

We chose this game because of the consistency of the emotional experience with clear timestamped markers of emotional events. Players do not have many opportunities to deviate from this fairly narrow emotional range, and playtime is long enough to be immersive. Although we might expect emotions to change if a player repeats a scene, repeated scenes provided an opportunity for multiple measurements of the same event.

Participants played through the introduction of the game, which took about 40 minutes, while wearing an EEG headset. A screen recording captured the gameplay, a camera captured their facial expressions, and the keyboard was outfitted with a force-sensitive resistor to capture keypress pressure information as well as keystroke timing. After a short break, participants reviewed the game footage with an interviewer. They were asked to characterize their experience in terms of a scale from stressed to relaxed. Particular points in the game were chosen to indicate the extrema of the scale, and then they were asked to characterize points in between with specific examples from their gameplay. Then, after annotating the footage with these characterized labels and ratings, they would review the footage once more, using a joystick to continuously label the timeline from stressed to relaxed.

Here is an itemized breakdown of the previous paragraph: (1) Gameplay is labeled with timestamps and qualitative explanations of emotional moments (2) These moments are used as concrete examples that make an emotion scale meaningful (3) The now-meaningful emotion scale is then used to label the entirety of the timeline

"Choose or Fuse" describes the above labelling process in detail. Data is available at https://www.cs.ubc.ca/labs/spin/FEEL_dataset. "FEELing (key)Pressed" reports on the initial results of the study. The practical upshot is that the pressure sensor data from keystrokes performed better than brain activity for predicting the emotion curve that was constructed by the labelling process above. Brain data was supposed to be the gold standard because that must be where the emotions "really exist." Yet, brain data underperformed. Why might that be?

My suggestion is that imagining emotional experiences as a signal processing paradigm is a fundamental misunderstanding of emotions. Let's break down the underlying assumptions.

Is emotion a signal? Clearly signals are involved in emotions. Electrochemical action potentials travel along neuronal chains which encode information from physical events. The signals are processed and distributed by

different neuron clusters in the spine and brain. There must be signals for there to be emotions. But the phenomenon that we call an emotion is not a single, independent "event." Even physical "events" are not processed independently in a complex system such as the body.

The brain is more like a distributed system than it is a single processor. An "experience" can be reasonably said to be an emergent property of neuron activation across different subsystems. In fact, it is not very clear that there is an "experience" part of the brain, as much a synchronization of distributed brain activations that happen to produce a subjective experience like attention (Ward [2016], Ward et al. [2006]).

For example, the James-Lange theory of emotion states that our emotions "are" only our cognitive reflections of body states (Lang [1994]). That is, what we experience as anger is simply the sensory feedback from tensed muscles, blood rushing to our face, sweat glands opening, etc.

Let's say that this is true: as physical phenomena, these are all measurable signals, in principle. Even so, the timescale of measurement would be non-uniform. Sweat glands have higher latency than muscle contractions (we'll be sweating long after our muscles stop tensing) (Lockhart [1972]). A single multi-variate slice of measurement of facial blood flow, muscle activation, and sweat presence gives us a body "state." Two or more slices can give us a trajectory vector for each signal (rate of change).

However, there is still a disconnect from our neuron activation as well as our subjective experience. Inattentively, certainly all of the information that we could externally observe would correspond to some portion of the nervous system in a state of activation. Subjectively, we very likely would not be able to attend to all of these signals simultaneously, or, even individually without fundamentally changing the subjective experience[6].

Another emotion theory, appraisal theory, reverses the causal lens, instead saying that it is our cognitive appraisal of perceptual events that *produces* the tensed muscles, activated sweat glands, and flushed face (Ellsworth [2013]). In that case, we would expect the brain activity to happen *before* the measured body responses.

The reality of the experience is that both directions of causality are continuously happening at many levels at the same time. The analytical mistake is in decomposing a fundamentally parallel systemic experience.

[6]Doesburg et al. document the subjective experience of visual attention as "flipping" between perceptual experiences if a research participant is given a different picture for each eye. That is, a participant will report seeing a butterfly, then a leaf, then a butterfly, repeatedly, back and forth. One could imagine that it would be a blurry "mix", but that is not what participants reported. (Doesburg et al. [2009])

The sum is greater than the parts.

The body and brain are a continuously acting system embedded within a local environment. The continual sense perception of the environment is always inflecting the signals themselves. One does not stop existing in-between stimuli; and, for a conscious cognitive system, to exist is to evolve.

For our collected EEG data, then we would have to ask: what exactly are we studying? Considering the fact that EEG data is only able to capture brain wave activity, rather than neuronal activation, and mostly at the level of the cortex (the layer of the brain closest to the skull), if we were to be collecting anything sensible, it would be the cognitive responses to prior stimuli. However, most of our emotion processing can be thought of as pre-cognitive (i.e., deep limbic brain activation), especially as it relates to other external responses such as muscle tension, blood flow, and sweating. Without an extremely high-fidelity functional map of the brain (which does not exist) and deep-brain data activation information (which EEG does not collect), it would be foolhardy to expect any kind of correlation between emotion data and brain data.

In fact, the EEG study illustrates the limitations of the approach of "throwing data at the problem." The fact that force pressure data collected on force-senstive resistors (FSRs) on the keyboard is a better predictor of emotion than brain data is likely because the pressure data is an externalized expression, much closer to something we might expect someone to attend to. The most advanced functional brain studies rely on extremely short (1-10ms) perceptual stimulation tasks and are only now mapping small portions of functional networks involved in perception. Trying to relate emotion data to EEG signals is like trying to map whale migrations with satellite imagery. You might see a whale here and there, but most of the stuff is going on under the surface.

Realizing the limitations of our most advanced sensing technologies is important. The next paper I present here outlines limitations and expands on the critiques from this introduction. The paper that follows was written with the recent EEG study in mind. Why did FSRs outperform EEG? My position is that we did not truly understand what we were trying to measure. This paper attempts to refine our idea of what we were trying to measure and why. The paper as presented is the full unaltered text from the journal article entitled "Affective robots need therapy."

4.1 Introduction

When a dog wags its tail, is it nervous or happy? The answer is likely either or both depending on the context. Did the dog's owner just come home? Is the dog's owner looking upset because they noticed a broken lamp in the living room? Does the dog usually get a treat when the owner comes home? Did the dog have a good day or a bad day? Let's say we wanted to build a dog happiness detection system into a robot. It would be difficult to ascertain ground truth, because dogs can't tell us how they feel. At best, we could ask people who know the dog pretty well to interpret the dog's behaviour and provide labels for a classifier. But, realistically, we would be better off building a dog-tail-wagging detection system because it would be grounded in observable quantities and just leave out the question of happiness.

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

If emotions are constructed, then we can think of an emotional experience not as caused by any singular portion of a robot's body or behaviour, but as emerging from the interaction as a whole, contingent on an interactor's narrative framing (Bucci et al. [2019, 2018], Marino et al. [2017]). Instead, the emotional meaning of specific behaviours is continually grounded (Jung [2017], Leahu and Sengers [2014]) through ongoing behaviours during the interaction. Recognizing that robots need to be programmed with structured, determinable data, we believe that the answer is not simply to give up on quantitative methods, but instead to embed them in constructivist philosophical approaches and methodologies. If you are studying objective phenomena, use objective methods. If you are studying subjective phenomena, use subjective methods. If you are studying both, use mixed methods.

Objective approaches are useful for studying phenomena that are objective, determinable, repeatable, and somewhat culturally-independent. But much of the emotional phenomena we wish to study within the field of affective robotics are not objective, and subjective approaches are more appropriate when interpretation is fundamental to the phenomena at hand.

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

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

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

1. a synthesis of the theoretical and pragmatic basis of therapeutic care methods and their meaning for affective robotics

2. an account of the constructed nature of emotions in HRI and errors that can result from not accounting for emotions-as-constructed in study design

3. resolutions to the above and accompanying methodological recommendations based in examples from therapeutic methods

4.2 Why emotions as constructed matters to HRI researchers

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

Intuitively, we can use an analogy of sports to understand why biophysics can be both culturally-embedded and highly-personalized: a weight-lifter will have a different body structure depending on their culture and personal preferences. Their body will depend on the people/places they interact with, e.g., their personal trainer will prefer certain exercises, the gym will have only certain equipment, their nutritionist will suggest specific supplements. Similarly, the weight-lifter's friends might value certain body shapes which will influence the weight-lifter's values about what to practice. And, on any specific day, the weight-lifter's immediate biophysical structure is contingent on other cultural and personal preference factors such as their breakfast, whether they stayed up late streaming T.V. shows, and so forth.

The experience of an emotion is only available to the subjectivity of the person experiencing it. Yet, in HRI, we rely on objective methods of measurement (e.g., sensors) and statistical methods (e.g., surveys) that treat emotions like they are universally experienced. Constructed emotions is a different understanding of the nature of emotions than is common in HRI. To study emotions from this ontological perspective requires a different epis-

temology. An *epistemic claim* is one about the way in which we come to know something, i.e., how we can study and produce knowledge about a phenomenon. In the next subsection, we articulate three epistemic approaches and their relevance for HRI.

4.2.1 Epistemology of Modern Science and Errors in HRI

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

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

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

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

We use a running example of studying "trust" via galvanic skin response (GSR) and provide four errors that are introduced into experimental methodology by avoiding the constructed nature of emotions (the authors themselves have made these errors numerous times). We use this example as a stand-in for HRI studies that take an emotional phenomenon (trust,

love, etc.) and purport to provide a causal link between that phenomenon and a signal (GSR, heart rate variability, robot pose, etc.).

1. **Categorical error**. Trust is better understood as an emotional construct or concept that includes a variety of contingent emotions rather than an emotion itself (Holth [2001], Simpson [2007]). By studying trust without making this distinction, the researcher makes a categorical error. The reason behind the categorical error is that trust emerges from cross-cutting ontological[7] and epistemic domains. That is, trust exists as a combination of somatic, behavioural, and cognitive aspects that are embedded in a cultural frame. In other words, our in-the-moment body feelings and senses, action and thoughts are constructed from a lifetime of experiences with other people. As a result, measuring trust is like measuring weight lifting — you can quantify aspects of weight lifting, but it makes little sense to ask people to rate 'weight lifting' on a scale of 1 to 5. A better design would ask participants to inspect the constituent emotions behind trust. One approach to resolving categorical errors is to *ground* (Jung [2017]) experimental terminology to ensure common understanding between researchers and participants.

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

Expecting participants in a study to introspect on their emotions without training will introduce uncontrolled and hidden variability. One way to account for this methodological error is including training into the study design. A sufficiently high sample size can also give insights to population-level trends, but also obscure individual experience.

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

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

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

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

4.3 Understanding the Constructed Nature of Emotions

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

4.3.1 Emotions happen all over the brain and body

We believe that having a good understanding how emotion happens in the brain and body can give a working mental model of the different kinds of emotional phenomena we attempt to study when designing robot bodies and behaviours. In particular, we believe it gives us a good understanding

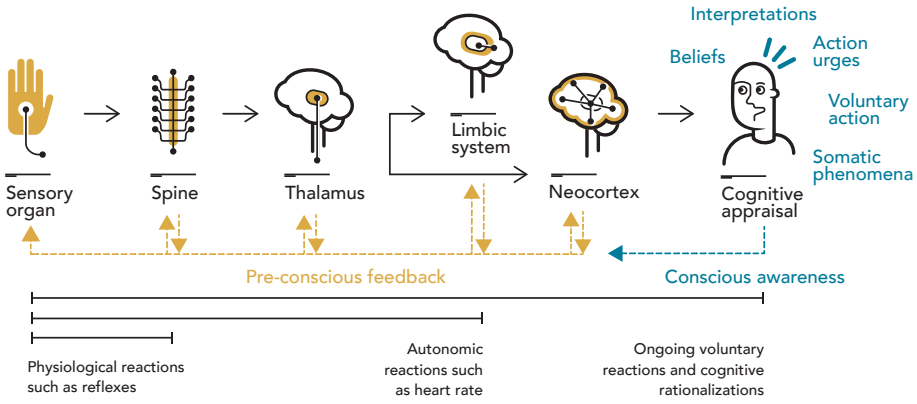


Figure 4.1: A spectrum of locations in the body where emotion may be said to occur. Rather than imagining the brain as a singular processing unit with top-down control, it is useful to think of different systems of the body acting at different timescales and with feedback into each other (Parent and Hazrati [1995]). An emotional event will be "experienced" by different parts of the brain differently, each of which is structured and "trained" differently (Sapolsky [2003]).

of why emotion experiences may be different between different people—an increasingly common viewpoint amongst emotion theorists. In fact, Ortony and Clore (of the OCC cognitive appraisal theory of emotion) present a summary of evidence against conceiving of emotions as universal experiences (Clore and Ortony [2013]):

> *Should one assume then that specific emotions do not exist? No, but perhaps some long-standing assumptions about them should be reexamined...emotions are not marked by distinctive behaviors or even by reliable patterns of feeling (Barrett, 2006)... Many assumed that affective neuroscience would rescue the study of emotion from this untidiness. However, a recent meta-analysis of imaging results concludes that the evidence that specific emotions have specific locations in the brain is not strong (Lindquist et al., 2012).*

In Figure 4.1, we illustrate a working map between emotional phenomena and the places in the body that they can be said to "happen." The emotional phenomena that we imagine as singular experiences have a biological basis in different parts of the body and brain. For example, let's

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

At some level, it is convenient to think of emotions as signals. Nerve signals are responsible for transmitting sensory information to the brain, for controlling muscles and other body parts, and, as far as we know, in some way actually comprise the conscious experience. If we trace a sensation starting from an external event (such as a sharp object activating a pain receptor), the signal would pass along nerve fibres to ganglion cells, to the spine, medulla, midbrain, thalamus, and then finally to the amygdala and somatosensory cortex. At each integration juncture, the signal may proliferate other signals; e.g., by instigating a reflex. However, the conscious experience of the signal would not be possible until the signal has been processed and distributed to the neocortex and amygdala through the thalamus. This means that bodily reactions are already occurring before we are fully conscious of sensation, and, further, that multiple parts of the brain will be processing the signals at different times. Much of what we think of as observable emotional signal are autonomic responses, such as increased galvanic skin response, heart or breathing rate. Even if we have some indirect voluntary control over these autonomic responses, the instan-

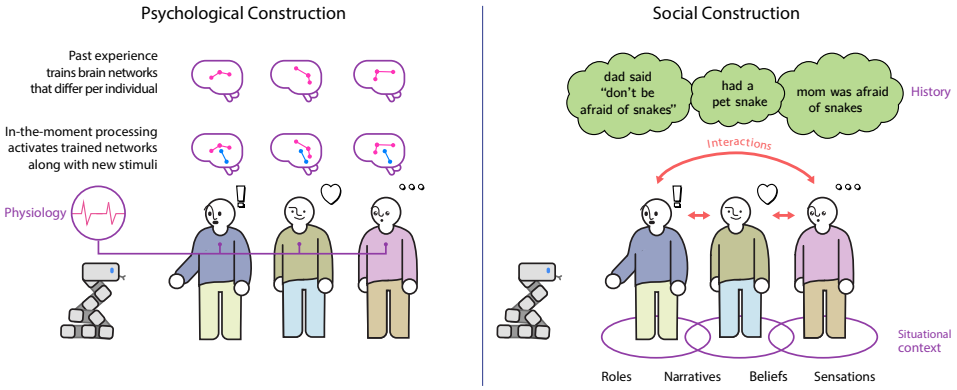


Figure 4.2: Emotion can be psychologically and socially constructed. We integrate past experience, narratives, social relations, and distributed brain networks in understanding our bodily sensations.

taneous reaction is not directly available to our conscious experience, rather, the post-hoc sensation of the autonomic response is available. That is to say, we cannot consciously choose to sweat, but we can notice that we have started to sweat after it begins.

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

4.3.2 Example of Emotions as Socially and Psychologically Constructed

As an example for the social and psychological construction of emotions, we extend an example from Barrett's paper on constructed emotions (see Figure 4.2 for accompanying drawing).

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

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

4.3.3 Evidence for emotions as socially and psychologically constructed

An implication of emotions as socially and psychologically constructed is that each individual's experience of emotional phenomena is highly-dependent on their own specific subjectivity, which is itself highly-dependent on their interactions with other people both in-the-moment and over a lifetime. Our subjectivity is (physically) constructed in the brain from a lifetime of experiences where we associate phenomena in the world (such as objects, envi-

ronments or other people's behaviours) with perceptions of the world and sensations in our body. To use a neuroscience-to-computer science analogy, we can think of construction as being both topological changes in brain networks as well as patterns of activation across brain networks (that also happen to reconfigure the network).

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

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

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

If we take this seriously, labelling emotional experiences with a singular word or point on a scale hides the unique and multifaceted experiential and physiological phenomena occurring during an emotion. This is not to say that we should not try to understand emotions and engage in scientific practices of labelling, categorization and structural modelling, but rather that we should approach emotions as socially and psychologically constructed and therefore fundamentally interpretive phenomena. That is, our conscious communication of emotion-related phenomena is necessarily dependent on interpretation and representation, based on incomplete introspection.

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

4.4 Therapeutic Approaches and How they Apply to HRI

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

4.4.1 Manualized therapies

Cognitive-behavioural therapy (CBT) is one of the most widespread therapeutic frameworks (Milne and Reiser [2017]). Emotional interventions are three constituent parts: behaviours, cognitions, and emotions (see Figure 4.3). For example, if a patient believes they are a "bad person," they may engage in behaviours that a "bad person" would do; then might feel

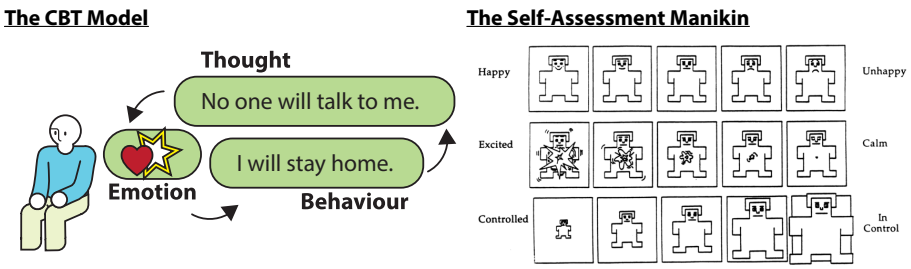


Figure 4.3: The CBT model (left) relates thoughts, feelings, and behaviours. The therapeutic concept is that you can intervene on any aspect of the cycle to change your emotions. The point of learning for HRI researchers is that therapists have developed a model such as this because it reflects a common and effective way for people to analyze and communicate about emotions. This is in contrast, e.g., to the self-assessment manikin (right) which only inspects an abstracted aspect of "feeling."

guilty; further reinforcing their original belief. CBT aims to intervene on this cycle by asking patients to identify: (1) their adverse beliefs (often by writing them down) then rehearsing a countervailing belief; (2) unwanted behaviours and rehearse alternative behaviours; and (3) unwanted emotions and rehearse alternative emotions. CBT has been effective at treating a wide variety of mental health difficulties and is heavily *manualized*, that is, CBT relies on manuals, workbooks and handouts (see Figure 4.6 for a DBT manual excerpt) to deliver both psychoeducational content and to help patients to practice CBT skills.

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

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

We mention CBT and DBT as they are common approaches, however many therapies have been manualized. Importantly, these have been developed and verified through practice and therefore both implicitly and explicitly include methods for grounding the meaning of the materials. For example, the DBT emotion worksheets help a patient to label their emotions by providing examples of possible somatic experiences, beliefs, behaviours, and contexts for an emotion. They do not expect a patient to understand the worksheets immediately: the patient works with the group and the coaches over many weeks and months to develop a subtle understanding of each emotion (see Figure 4.6).

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

4.4.2 Somatic therapies

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

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

Rather than assuming that an emotion is easy to identify and label, the fundamental assumption of somatic therapy is that many different metaphors

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

4.4.3 Narrative therapies

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

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

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

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

Particularly for studies in which a robot is presented as a social actor, the narrative that the participant develops about the robot can entirely deter-

mine their emotional perception of the robot. This is evident in a number of HRI studies (Jung and Hinds [2018], Marino et al. [2017], Bucci et al. [2018], Ling and Bjorling [2020]) that have studied narrative's impact on emotional ratings of robots. Even if HRI researchers would prefer to ignore the interpretive elements of narrative interaction, it is obvious that participants will engage in narrative interpretation whether or not the researchers would like them to.

4.4.4 Trauma-informed approaches

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

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

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

A critical look at the institutional relations of the hospital would point

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

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

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

4.5 Accounting for Subjectivity in HRI Study Designs

In this section, we (1) conceptually address the four errors outlined in Section 2.1; and (2) provide worked examples of how we imagine HRI researchers could work with therapists to solve those errors. We provide illustrated examples to explain our position in the text, and have included worksheets in Figure 4.6.

4.5.1 Addressing (1) categorical errors.

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

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

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

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

However, if we view stress as constructed, we would have to account for the ongoing simultaneous but categorically different factors that comprise the stress experience. We would want to account for and differentiate the

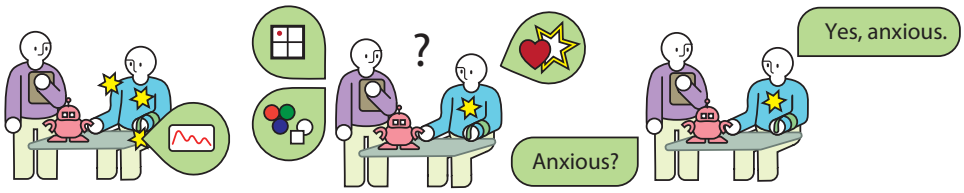


Figure 4.4: In this figure, we see the researcher (purple) and participant (blue) engaged in a galvanic skin response (GSR) trial related to stress. The yellow stars indicate that the participant's many somatic experiences, only one of which may correspond to the GSR graph. Further, without discussion, we do not know what the participant is experiencing as their stress response, or which aspects of the stress response are being captured in the GSR. Clarifying this would help us understand the categorical differences between stress-as-measured vs. stress as experienced. Then, we see the researcher asking the participant to represent their experience with a common 2D affect grid and using shape and colour metaphors to expand their shared understanding. Last, we see the participant and researcher agree to call that somatic experience "anxious" which can serve as a grounded term for the rest of the study.

participant's immediate evolving somatic experience from their cognitive rationalizations, which means accepting that a participant can only communicate with limited available language. If we had to capture a "stress level," we would commit to spending time with the participant to substantiate which of the categorically different parts of the stress experience we would like them to introspect about.

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

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

Somatic examination may reveal that stress was phenomenologically different enough between participants to be a meaningfully different kind of

emotional response, which researchers would want to account for in post-hoc analysis.

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

4.5.2 Addressing (2) methodological errors.

In contrast to how we expect our study participants to be able to make on-the-spot emotional assessments, therapists usually train clients over long periods of time to introspect and determine a variety of emotional phenomena. In an interaction, there is a subjective interplay between beliefs, behaviours and bodily sensations. Each therapy assumes that introspection requires a therapist to train and practice with their clients to determine a variety of emotional phenomena (see Table 4.1). This is in contrast to implicit assumptions in HRI studies that participants should be able to "dead-reckon" their emotions without much training. We argue that HRI researchers should form their methodologies with the principle that *emotions are difficult to introspect accurately*.

Methodological errors occur when this principle is violated. However, it is understandable that it would be violated because of practices within academic psychology that reasonably try to limit the impact of researcher bias. For example, it is common to treat participants as "blank slates" and provide "validated" surveys and treatments as if they are neutral experimental factors. For example, the Positive And Negative Affect Schedule (PANAS) is a "validated" mapping of emotion words to affect grid quadrants. Or libraries that map movie clips to emotion ratings (Gabert-Quillen et al. [2015]); the assumption being that it is a standard treatment factor that can be applied to produce a particular emotion in a participant.

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

A constructed view of emotion would entail that we expect each participant's experience of an emotional phenomenon to be different. Further, we would imagine that their reaction would be sensitive to conditions. As such, we would not know whether these validated emotion scales and factors are,

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

Table 4.1: A list of phenomena of interest that comprise an emotional event inspired by a framework from dialectical behavioural therapy (Dimeff and Linehan [2001]). Objective phenomena (O) have causally observable quantities that can be measured and compared using physical sensing equipment (sensors, rulers, etc.). Interpretive phenomena (I) require some adjudication through language and introspection. Behaviours and expressions are differentiated here by duration and level of attention, i.e., a behaviour requires at least some attentive voluntary control, but expressions may or may not require attentive control. The objects and actions within events are observable, but categorization requires some interpretation. See Appendix for more examples.

in fact, producing or capturing the subjective experience we expect. We would have to rely on the quality of participant introspection to trust our measurements.

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

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

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

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

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

Drawbacks. The above process would add time to the study and require training for the researcher. However, the stronger critique is that it introduces researcher bias into the study. This could become problematic in larger-N studies with many different research assistants who run participants, presenting a greater need for consistency controls. The process also

excludes non-in-lab surveys as a possible method since it requires iterative feedback.

4.5.3 Addressing (3) instrumental errors.

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

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

This has precedence in pain management (Rosier et al. [2002]). Pain is understood as a highly personal experience: someone's previous experiences of pain affect their current experience of pain, and cognitive beliefs relating to their pain are known to impact the emotional processing of that pain (Lamé et al. [2005]). As such, doctors will administer pain measurement scales in a way that benchmarks the scale by asking the person to imagine the most and least pain that they have experienced to ground the meaning of a '10' and a '0' (Ong and Seymour [2004]). Studies in symptomatology incorporate metaphors to help a patient describe the experience of their pain (e.g., a sharp pain or a throbbing pain) which can aid in diagnosis or therapeutic reconceptualization (Gallagher et al. [2013]). Studies that attempt to aggregate pain measurements across patients have to account for this individual variability (Manworren and Stinson [2016]). Further, it is understood that the act of measuring can often heavily influence the outcome of the scale measurement. For example, one study showed a large discrepancy between the amount of pain patients reported on paper scales administered by nurses in person vs. electronic scales administered remotely (Price et al.

[2018]). The important lesson with scales for pain management is that even if we assume some universal mechanism for sensing pain, the perceptual aspect may be significantly different due to different past experiences that our brain was exposed to. Further, how we express and describe pain is influenced by our beliefs, ability to remember past pain, the social dynamics of the measurement process, and our understanding of the meaning of the scale.

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

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

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

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

For example, it is an ongoing debate within quantitative psychology as to the validity of treating Likert scales as continuous linear variables (Pimentel and Pimentel [2019]). We can understand why HRI researchers would prefer to treat them that way, particularly for regressions. It also complicates between-participant analysis: if one person's baseline is different than an-

other's, or one person's metaphor is different than another's, can we validly group them during post-hoc tests? Incorporating metaphors into analysis could therefore decrease statistical power by virtue of having more blocks, factors, or groups.

4.5.4 Addressing (4) Complexity errors.

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

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

Taking a feminist narrative therapeutic perspective would suggest looking at emotional interactions from a critical narrative lens by situating the participant and robot relative to the participant's self-understanding and perceived power dynamics. Dramaturgical theories of emotion align with certain critical feminist perspectives, as emotional expressions are assumed to be fundamentally performative. A study from this perspective would examine the conflict between a robot's intended displayed emotion, actually-communicated emotion, and internally-felt emotions. Behaviour from this perspective is thought to be representational of internal states, but abstracted and mediated through identity and social norms[8]. In the view of dramaturgical emotion theory, what is expressed during the interaction has a different emotional tenor than the subjective experience of each person alone. In affective robotics terms, assigning an emotional label to the behaviours would not give the only reading of someone's internal emotional experience, but instead what they felt it was appropriate to convey (Turner and Stets [2006]). Symbolic interactionist theories centre the reinforcement of one's

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

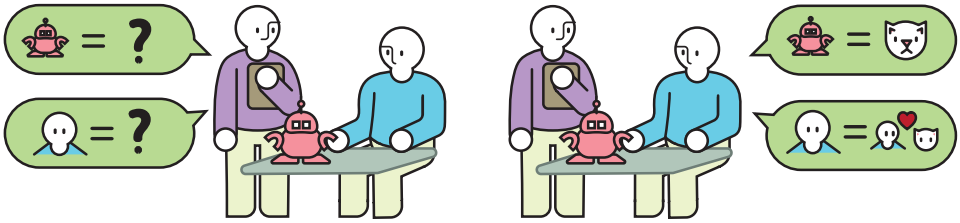


Figure 4.5: In this figure, we see the researcher speaking with the participant about how they see the robot within a narrative. By grounding the robot with a narrative such as "the robot is like my cat," both the participant and the researcher have established the boundaries of the "suspension of disbelief" that is necessary to see a robot as an autonomous agent. It further explicates the participant's emotion reactions within an accessible cultural frame. The emotional relationship between cats and owners are known as cultural concepts which provides common referents.

own self as the primary objective for emotional motivation, where identity may include multiple, overlapping identities (Stets and Turner [2014]). Symbolic interactionists imagine emotion as a continuous process that produces and also results from identity. Identity is continually negotiated with regards to cultural norms, beliefs, and social roles.

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

An HRI approach that uses therapeutic practices founded in social science theory can help to address complexity errors because of the awareness that social theory can bring to often-unseen cultural forces. They can help to expose hidden variables, provide language for roles/responsibilities/beliefs that are impacting a participant's emotions, or serve as a theoretical basis for analysis. In the previous example, the therapist is able to use a feminist framework to go beyond a surface level understanding of their patient's emotional experience by exploring the sociological factors that shape it.

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

Example. During a study, an HRI researcher would try to address which kinds of social dynamics might be at play. The robot's "story" can be provided by the researcher, or built with the participant. The social role of a robot can drastically change participant perceptions of emotional behaviours (Chen et al. [2020]). Whether or not we intend it, robots can be seen by participants as existing in a make-believe world (See Figure 4.5). Narrative therapies and interpretive emotion theories would provide insight into how to help resolve this.

A *feminist narrative perspective* would encourage critical reflection as to how robots are integrated into systems of power. For example, a teaching robot would be examined in terms of the role of a teacher in producing emotions, as opposed to the effectiveness of administering information (see Figure 4.5 for a illustration of roles).

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

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

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

to mix theoretical approaches due to apparent philosophical incompatibilities.

4.6 Discussion

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

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

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

emotion understanding in their daily work.

4.7 Conclusion

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

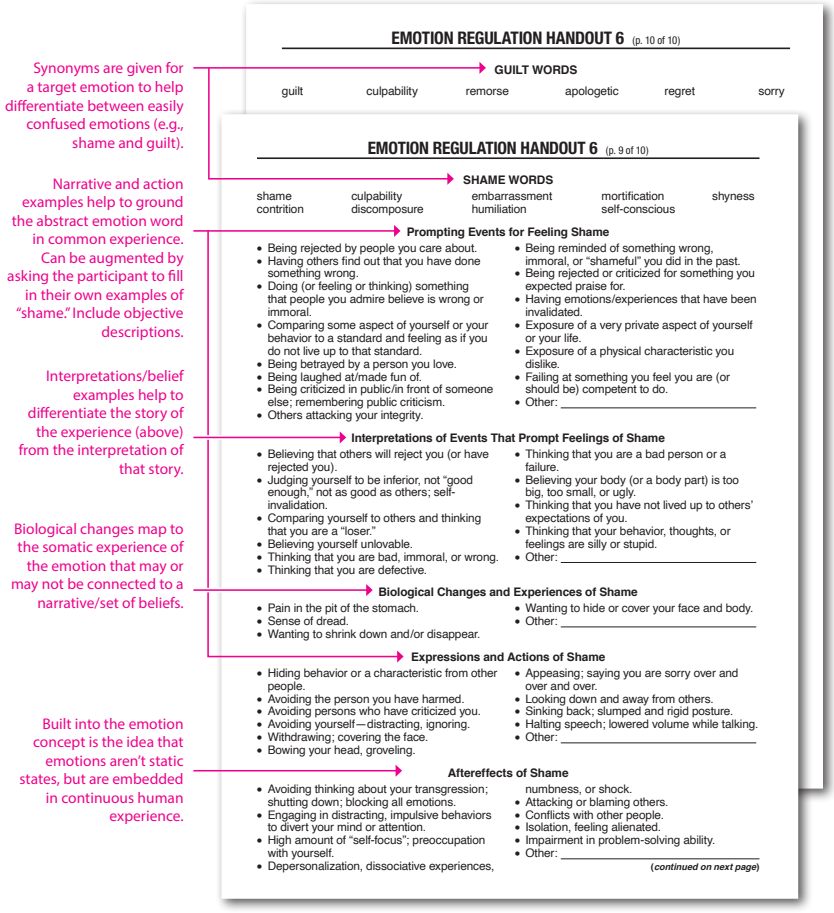


Figure 4.6: An excerpt from the DBT manual (Linehan [2014]) for emotion words. This can be directly used by HRI researchers if shame is meant to be studies, or adapted for emotions of interest. Provided by contrast for guilt (above) which are emotions that are commonly confused and may be useful to differentiate. Emotion sheets such as these can give context to emotions and can help to ensure that participants have grounded concepts using which they can differentiate their self-measurements.

Chapter 5

Teleoscope Systems Paper

A summary of the critique of the previous sections is that emotions are constructed through a system of meaning that interacts between the cognitive and sensory parts of a person's body and brain, which themselves are embedded and trained by a social context. What we can label as meanings, cognitions, and emotions can shift very quickly depending on context.

Computer scientists are quite familiar with this through the lessons of machine learning and neural networks. Neural networks are typically very sensitive to initial conditions, hard to predict, etc. This phenomenon is well-known by people who study adversarial attacks in machine learning where well-tuned noise can be added to an image to radically alter classifier predictions. Famously, images of cats can be made to be mislabeled as bowls of guacamole with small enough amounts of noise that a human could not tell the difference (Athalye et al. [2017]).

Real neural networks are not so different, particularly when emotions are involved. We are not machine learning classifiers, but I would argue that it stands to reason that we are *at least* as complex. Therefore, looking for simple formulas, rules, and even small machine learning models is to misunderstand the formal structure of the phenomenon that we are investigating. Top natural language processing (NLP) machine learning models such as large language models (LLMs) have billions of parameters to be effective at simulating anything close to meaningful human writing (and I would suggest it is still a long way off).

LLMs may be a good formal model for understanding the specificity and dynamics of meaning. If we think about an arbitrary symbol such as a word like "cat," there are a large variety of associations that it might evoke. Prior to LLMs, common NLP approaches to word senses were to create trees that encoded word categories. For example, "cat" would be under the "animal" category and similarity to "dog" would be defined in terms of number of steps to the most recent common ancestor. However, when I think of "cat" there are many other words that I might associate with the sense memories that come up, such as furry, fuzzy, sharp, food, litter, etc. Although many legions of computer scientists attempted to model these relationships with

explicit formal logic relations, the statistical approach of machine learning simply encoding association won out as more effective.

So, with an LLM, the "meaning" of cat is actually a fairly complex structure of associations between different symbols. In the high-dimensional space of vector representations of these symbols, the initial conditions may radically change the high-dimensional location that has been "activated" by a prompt. Learning from LLMs, we can see that there is some truth to the behaviourist claim that we are "just" association machines. Symbols themselves are emergent from sense data.

John Searle talks about the phenomenological difference between experiencing a direct perception and the memory of a perception (Searle [2015]). Hearing a sound is never exactly the same as remembering a sound, even if the person doing the remembering has a very good auditory memory. That is because memory is more reconstruction than recall. We operate more similarly to LLMs in our memory systems than classic computer file retrieval. A memory is not encoded directly as the same set of signals that entered the brain, nor when it is retrieved do the exact same parts of the brain activate. There is no particular location for a single memory; instead the memory is encoded across the brain as a whole.

The complexity of the whole cognitive system can be illustrated by the difficulty in trying to remember song lyrics. Sometimes, it is required to go through the entire song to remember a particular lyric. Sometimes, it is required to whistle a few bars. The memory is contingent not just on a static encoding within the brain from a single area, or else we would be able to simply recall directly. Instead, the system needs to be in a particular state to achieve the reconstruction of the memory.

5.0.1 Implications: Design for Externalization

For me, being able to understand meaning systems in this way has very strong and important design implications. That is because we must think about meaning systems as being both internal and external to the body if meaning systems are so sensitive to initial conditions. That is, there is a true sense in which a meaning does not actually exist in the person as much as the system of the person and their world. Designing for that reality requires a focus on *externalization* and *extending* our cognition into the world.

If we use a whiteboard to work out algebra, is it not part of our cognitive system? The whiteboard is clearly not part of our brain, but it is required to perform the calculation. In fact, the entire bodily procedure can be said to be part of the calculation, the physical moving of our hands, seeing the

writing appear, the complex continual interplay between perception and action. The end of the procedure gives us an answer which we consider to be the outcome of the calculation, but a full account of the process includes body, brain, and whiteboard.

This is an example of *extended* cognition. The whiteboard extends our cognitive system and involves action and human sensation in the process of cognition. At first, this can seem to challenge commonly held ideas about cognition, i.e., that it all happens "in the brain" through "thinking." But talking about extended cognition allows us to answer questions about why cognitive systems work they way they do. Particularly as designers, this helps to give an explanatory theory to why it even matters to make interfaces.

My contention here is that a meaning system is like a cognitive system, but includes further extent spatially and temporally. Specific symbols "contain" meaning not because the information content is actually contained in the symbol, but because the symbol, like a puzzle piece, is given meaning through a context of a person's cognitive system, environment, and culture. This is why something can be represented with a low number of bits, but have larger information content within the system that processes the symbol. Designers know this intuitively; however, it explains why some symbols are successful and some are not in design.

Taking the lessons of extended cognition seriously has deep implications for design work. For example, we should expect systems to be able to only develop meaning through externalized interactions, and for those meanings to be revealed only through a process.

5.0.2 Designing Teleoscope

The next part of this dissertation transitions from affective computing to creating meaning while interacting with data. I personally think of this as taking the lessons I learned through studying meaning as an affective phenomenon and applying them as a designer to a difficult and interesting problem. If meaning shifts so easily, then a system that helps a person make meaning out of their data should reflect that by helping manage the complexity, create predictability out of the chaos, and continually provide external cognitive markers.

Teleoscope comes out of a simple idea: what if we made a computational whiteboarding system so that the spatial associations on the whiteboard refined a machine learning model of a large dataset? Essentially, it was a design project that tried to get an externalized system to reflect an inter-

nal meaning structure. By understanding meaning structures as chaotic, we would understand that the process of interacting with data would change the meaning of the data for a researcher. By understanding meaning structures as complex, we would understand that many cross-cutting dimensions of association would need to be simultaneously represented, but with limited screen space. By understanding meaning as holographic, however, that allowed us to offload the meaning representation into a process (rather than a single symbol) and continually remind researchers of "how they got there" when coming up with a particular meaningful interpretation of our data.

What follows in this chapter is the paper that was submitted to UIST which describes the three-year process of designing Teleoscope. The next chapter will include our understanding of the implications for qualitative methodologies for large datasets.

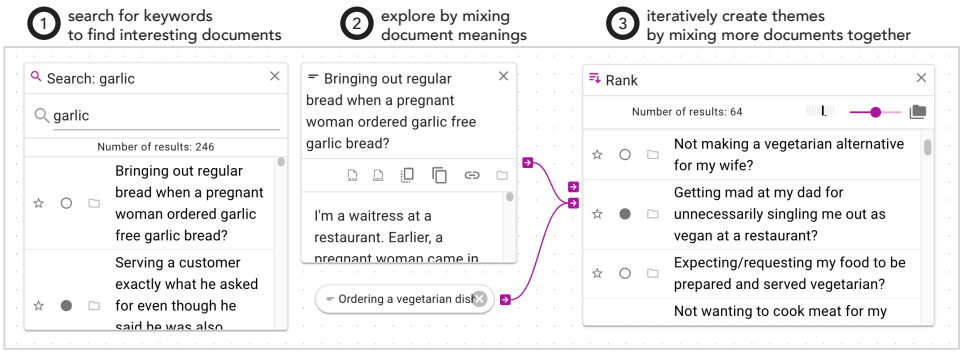


Figure 5.1: A screenshot of the core Teleoscope workflow: starting from a keyword search, you choose documents to iteratively mix together, and then explore the ranked output documents.

5.1 Introduction

Exploring data at scale within a large corpus of documents is difficult, particularly if you want to interpret your data by telling a story or explaining a phenomenon. Qualitative research often focuses on interpretation, which we can think of as enriching data with context and meaning (Hennink et al. [2017], Sebele-Mpofu [2020], LaDonna et al. [2021]). For large corpora, quantitative statistical approaches are often used at the cost of qualitative interpretation. Interpretation is often done by hand, which involves extensive reading and annotating data while developing themes during a *thematic*

analysis (Braun and Clarke [2012]). But, working by hand is slow and limits the scale of what can be analyzed.

For large document corpora (100K to 1M), some machine learning (ML) support is necessary due to the large volume of data. With the advent of large language models (LLMs), modeling context and meaning is becoming possible, providing the potential to automate interpretation. However, a rigorous qualitative researcher cannot simply input data into ChatGPT and ask for a summary. Instead, researchers need to demonstrate the validity of their interpretations by showing the *provenance* of their analysis—to themselves, to their colleagues, and to reviewers. When analyzing, researchers aim to produce cross-cutting *themes* by drawing out related meanings between documents using interpretation and theory.

In the context of LLMS, "relatedness" is encoded as distance between vectors that represent documents (a.k.a., document embeddings) (Devlin et al. [2018], Brown et al. [2020], Touvron et al. [2023], ChatGPT, OpenAI [2023]). To help create *themes*, a LLM would need to represent vector similarity in a way that corresponds to how a researcher thinks documents are related. From a computational perspective, this means finding a way to represent underlying meanings and subtext using document embeddings.

Creating these vector representations of meaning and subtext is no small task, but particularly while a researcher is actively trying to make sense of their data. In this paper, we are calling the "way we think" when making sense of data a *schema* (Berret and Munzner [2022]), and support the *thematic exploration* of data (Li et al. [2023]) by creating a system that helps to track how a researcher thinks about their data, i.e., models their schema. However, schemas are hidden, unique, and change every time the user interacts with data. Further, qualitative researchers often do not work alone; it is necessary to inspect, retrace, and share your thinking processes with other people on a research team.

To enable collaborative human and machine sense-making, we built Teleoscope, a web-based system to support multiple users interactively exploring a large corpus of documents such as social media posts (100K-1M) based on LLM document similarity. Teleoscope assists qualitative researchers during the discovery, data curation, and organization phases of a research study. To understand one's own thinking process, for purposes of rigor and sanity, researchers need to continually justify "how they got here," not just to colleagues, but to themselves. We therefore focused our design efforts on visualizing data *provenance* to support rigorous and opinionated refining of a dataset, which we call *thematic exploration*.

To design and evaluate our system, we ran three field deployments. In the

first deployment (N=5), we asked qualitative researchers to use Teleoscope to explore a shared topic over the course of a few weeks, then discuss their experience and findings at a focus group. For the second deployment, we recruited a three-person qualitative research team of nurses to use Teleoscope as part of their data curation and analysis process and responded to their design requests. For the third deployment, we released a live public version of Teleoscope which is currently running, available, and has active users who provide feature requests.

Our studies demonstrate that Teleoscope supports professional researchers in general-purpose text data exploration and discovery. We performed a software design process over two and a half years which was heavily informed by our own use of Teleoscope for qualitative analysis. To summarize, our contributions are:

- **C1. The design of a system that supports thematic exploration**. Teleoscope enables an extension of thematic analysis into document curation to enable more rigorous data interpretation which we call *thematic exploration*.
- **C2. Results of case studies with qualitative researchers.** Our long-term field deployments of the system demonstrate that qualitative methodologies can be supported with our approach of schema externalization.
- **C3. An open-source, usable, and adaptable product that supports collaboration.** Teleoscope is provided as a research product that is publicly deployed at [anonymized for review] and is open-source software. Multiple users collaborate in exploring over 500K documents on the same workspace. The system is cloud-native, works in the browser, and is robust enough to run continuously.

5.2 Design Process and Background

In this section, we outline our design goals and the relationship to existing literature. We explain relevant technologies, outline qualitative research theory, and motivate design decisions. We situate Teleoscope as a system for externalizing researcher's internal schemas about their datasets.

The concept of a *schema* comes from psychology and refers to a cognitive framework that helps to organize and interpret particular events, behaviour, and information. For this paper, the intuition, curiosity, and ad hoc hypotheses that researchers have about their data can be understood as

expressions of their schemas. An important part of the qualitative research process can be thought of as identifying, expressing and externalizing the researcher's schema through labeling, arranging, modifying, and enriching data (Berret and Munzner [2022]).

The word "Teleoscope" is a portmanteau of telescope and the Greek word *teleology*, which is about explaining the meaning and purpose of something. Teleoscope is a way for researchers to intuitively explore and explain meaning and purpose for their datasets. We use the natural interactions of a qualitative researcher when interpreting documents such as *arranging*, *annotating*, and *grouping*.

Our design goals target continuous live interaction, as opposed to topic modelling packages that run as terminal-based Python programs (Pedregosa et al. [2011], Rehurek and Sojka [2011], Devlin et al. [2019], El-Assady et al. [2019]). We visualize the *process* of thematic data exploration so that collaborators can share their thought processes and conclusions. This is different than topic modelling visualization systems that focus on high-level statistics, topic hierarchies, word counts, and labels (El-Assady et al. [2019], Kim et al. [2019], Terragni et al. [2021], Miner et al. [2023]). To support our target users—who are not computer experts—we created a web platform to facilitate fast enough interaction that users could feel as if they were having a creative and improvisational experience with the data.

Our research design goals are:

- **DG1. Imbue common qualitative research interactions with computation.** Qualitative researchers are familiar with *arranging*, *annotating*, and *grouping* of documents and use emergent interpretive processes such as physically re-arranging and annotating documents until patterns and meanings emerge.

- **DG2. Help researchers externalize their schemas as visual traces that can be *retraced*, *modified*, and *inspected* by collaborators.** When interpreting data, capturing the steps of a thought process is useful in collaborating with team members to inspect, share and modify workflows and understandings.

- **DG3. Facilitate an on-going feeling of improvisational, creative, live data exploration for researchers.** The feeling of continual interactivity was important to us to facilitate through our low-level system design choices, not having users be creatively blocked by waiting for long computations.

5.2.1 What qualitative researchers need

Teleoscope's design is largely inspired by the qualitative practice and methodological values of thematic analysis (Braun and Clarke [2012]). This is well captured in the following quote by Nowell et. al:

> *To be accepted as trustworthy, qualitative researchers must demonstrate that data analysis has been conducted in a precise, consistent, and exhaustive manner through recording, systematizing, and disclosing the methods of analysis with enough detail to enable the reader to determine whether the process is credible. (Nowell et al. [2017])*

Teleoscope takes the same value propositions as a thematic *analysis* phase and applies it to the *data curation* phase, i.e., collected with an interpretive perspective (Mannheimer [2021]). Data curation must satisfy the thematic analysis standards of being rich, precise, consistent, and exhaustive. Information power (an evolution of theoretical saturation) is used in qualitative analysis as an analogue to statistical power in quantitative analysis (Malterud et al. [2016], Guest et al. [2020]). At scale, there may be hundreds or thousands of documents that are thematically "identical" and provide no extra information power relative to their research question. Researchers therefore need to determine which documents are most similar within a corpus, structure the documents into groups, then find exemplar documents for more in-depth analysis. Teleoscope supports data curation by providing potential theoretical rigour that would apply standard thematic analysis by hand by focusing on visualizing the *provenance* of the ML model.

How then do we decide which documents are worth looking at? In quantitative methodologies, the standard approach is that scientists formulate their research as a series of falsifiable claims which are supported or rejected due to empirical evidence. In qualitative methodologies, a researcher's job is to *interpret* evidence so that it makes sense for both the researcher and the reader. Berret and Munzner explain the process of sense-making as moving between tacit schemas (i.e., gut feelings, unconscious ideas) and explicit schemas (i.e., drawings, writings, and other data representations) (Berret and Munzner [2022]). By turning a tacit schema into explicit schemas, a researcher makes their tacit schema available for inspection and critique.

Importantly, sense-making is an iterative process that starts from ambiguity. Chen et al. (Chen et al. [2018a]) argue that ML support for qualitative analysis in social science research needs to "*[shift] the focus to ambiguity.*" To paraphrase, computer scientists often create supposedly definitive and

"accurate" ML models of textual data too early, before there has been a period of ambiguity. Creating models too early is antithetical to qualitative definitions of academic rigour where it is considered important to preserve ambiguity for as long as possible. Having more time to make connections between data points can enrich theoretical insight while the researcher develops themes about the data. Good themes are *rich* (more interconnected) rather than *accurate* (categorically definitive) (Moser and Korstjens [2017]).

With Teleoscope, the act of arranging documents takes place within a node-based visual workflow editor. Arranging and linking elements to make a schema explicit updates the ML workflow as the user "updates" their implicit cognitive schemas. Braun and Clarke emphasize that during the data familiarization phase, there needs to be *active* engagement with the data:

> *Note-making helps you start to read the data as data. Reading data as data means not simply absorbing the surface meaning of the words on the page, as you might read a novel or magazine, but reading the words actively, analytically, and critically, and starting to think about what the data mean. (Braun and Clarke [2012])*

We consider Teleoscope to be a method of capturing the process of engaging with data that creates a correspondence between the externalized annotations on the screen, computational meaning with an ML workflow, and internal meaning for a researcher.

Teleoscope contributes to qualitative methodologies by importing standards of analysis data curation at scale (**DG3**). It also demonstrates a method for making the process of tacit sense-making explicit by asking the user to primarily manipulate workflows which represent their sense-making process (**DG2**). Unlike other systems, Teleoscope makes a schema expression an explicit goal, and provides a direct correspondence between the schema as expressed through visual workflow and ML model (**DG1**).

5.2.2 Visualization Approach

Our focus on visualizing provenance comes from a desire to make an ML system that can be used by qualitative researchers. These people are not computer experts, yet still need to explain their use of an ML system to stakeholders, such as collaborators, supervisors, and clients (Arrieta et al.

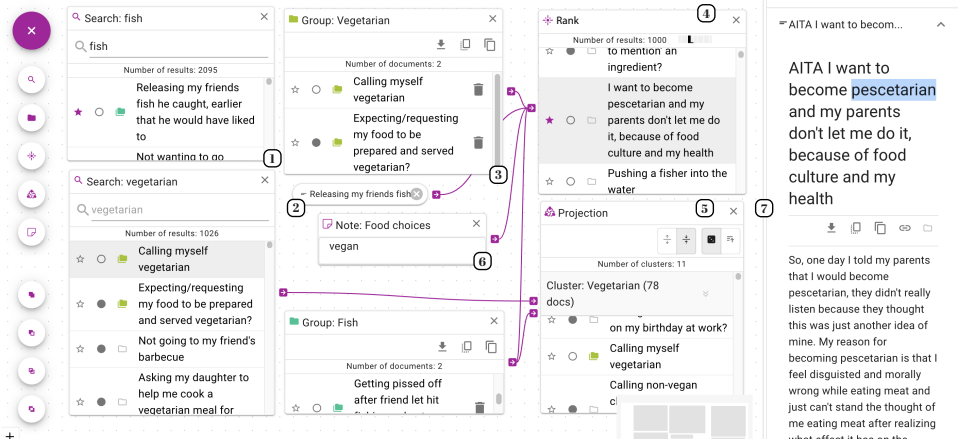


Figure 5.2: An image of the Teleoscope workspace. (1) Users start by performing a **keyword search** to explore documents; (2) **Documents** are dragged onto the workspace; (3) Documents can be put into **groups** for organization; (4) **Rank** nodes can use documents, notes or groups as control inputs; (5) **Projections** create clusters using groups as control input; (6) **Notes** can contain arbitrary text which is also vectorized and can be used as a control input to a Rank; (7) the **sidebar** has a quick viewer for documents, saved items, bookmarks, and settings. Keyboard navigation is used for quick exploration and group creation.

[2020]) (**DG1**, **DG2**). Many topic model visualizations choose to directly visualize the underlying document clusters, that is, the *result* of the clustering process (Nikolenko et al. [2017], Rüdiger et al. [2022]). In contrast, Teleoscope focuses on capturing and displaying the exploratory *process*. In the systems and visualization communities (Xu et al. [2020], Yuan et al. [2021]), this is referred to as *provenance* tracking. For making sense of data, the importance is placed on creating a visual trace of a history of user's interactions with the data to answer not just *"What are the results?"* but also *"How did we get here?"* (Hearst and Degler [2013]). A standard provenance system that we are all familiar with is *undo/redo*; a visualization of undo/redo could be as simple as a list of previous actions performed on the interface.

Beyond simply keeping an undo/redo history, researchers are very interested in creating reproducible and explainable results (Silva et al. [2007], Xu et al. [2020]). This helps during the process of analysis, publication, and review. For Teleoscope, keeping track of the provenance of an ML result

means tracking the inputs and user-selected data processing elements in a chained workflow. The user is engaged in a process of data wrangling for sense-making (Bors et al. [2019], Badi et al. [2006]), i.e., *exploration*, *annotation* and *curation* of data (Munzner [2014]). In *Trrack* (Cutler et al. [2020]), Cutler et al. demonstrate a library for tracking branching histories for actions on web-based visualizations. This is a maximal approach for provenance, where every action is tracked and is reproducible. Histories can be linear, branching, or networked, allowing for both high-scale overviews and detail-on-demand approaches (Yuan et al. [2021]).

Although our Teleoscope system does record every user action in a comprehensive history system, we hide most of that history from the user. Instead, we rely on the user creating small, comprehensible workflows that they commit to via grouping and connecting in a workflow graph. In Xu et. al's taxonomy of topic model approaches (Xu et al. [2020]), we are modeling coupled user and application state via an entity graph using semantic interactive visual analysis.

Making document sources the primary manipulation objects in a workflow was not our first design choice nor is it often what visualization systems choose. Similar to other systems (Lee et al. [2021], Terragni et al. [2021], Gad et al. [2015], Choo et al. [2013], Kim et al. [2020]), we first tried keywords and topic lists as the primary manipulation objects using a dashboard metaphor. However, we arrived at design insights that (1) the process of iteratively arranging documents itself is critical to sense-making; and (2) the spatial arrangement of documents often captures their relationships, which are important for remembering and inspecting thought processes.

Teleoscope applies ideas from provenance research to qualitative understanding of corpora. We chose to make provenance traces our primary manipulation objects. Therefore, researchers are directly creating the story of "how we got there" in a visual form that can be shared with collaborators, inspected, and updated (**DG1**, **DG2**).

5.2.3 NLP Approach

To facilitate the concept of *themes* over *topics*, we take the approach of *modelling by example* to allow users to explore and then structure the document space. Lissanddri et al. (Lissandrini et al. [2019]) describe example-based search approaches as having a resurgence in popularity. A variety of example-based query systems have been introduced that attempt to synthesize the query from examples (Martins [2019], Fariha and Meliou [2019],

Lissandrini et al. [2019]). Teleoscope follows these works by using semantic vector similarity to drive exploration. Teleoscope differs from these works by not attempting to construct deterministic queries or models of the entire corpus. Instead it relies on the interaction process and the user's own sense of information power to determine the extent of the exploration.

When a Teleoscope user has finished a phase of exploring via semantic similarity, they can switch to structuring the document space via semi-supervised dimensionality reduction. Variations on this approach are used in recent human-centred ML and visualization systems (El-Assady et al. [2019], Meinecke et al. [2021], Asudani et al. [2023], Sperrle et al. [2021]). The premise is to take a large language model and reduce dimensions along which a similarity metric is defined. Teleoscope uses the Universal Sentence Encoder (USE) for the base exogenous model, which encodes all documents as 512-vectors (Cer et al. [2018]). In Teleoscope, we use grouped documents as control inputs to define the similarity metric (i.e., this is the "supervision" part of our semi-supervised topic model approach). We use Universal Manifold Approximation and Projection (UMAP) for reduction to five dimensions (McInnes et al. [2018]). Teleoscope clustering uses Hierarchical Density-Based clustering (HDBSCAN*) (McInnes et al. [2017]).

Teleoscope uses a typical assembly of already-existing NLP tools, but uniquely capitalizes on a conceptual and practical correspondence between semantic example-based search, dimensionality reduction, and cognitive schemas (**DG1**).

5.3 System Design

In this section, we outline our design decisions for Teleoscope. Our design process took over two years and included a variety of design iterations, deployments, and studies. We report on the studies in the next section; this section outlines the design of our current live public release.

5.3.1 Teleoscope interface concepts

Teleoscope consists of an infinite whiteboard *workspace* where users arrange documents and operations into workflows to create a curated dataset. There is also a sidebar for quick navigation and reading. Users create workflows by making chains of sources and targets that control ML operations. The input of a Teleoscope workflow is a set of documents. The output of a Teleoscope workflow is a set of documents that are organized by thematic groups.

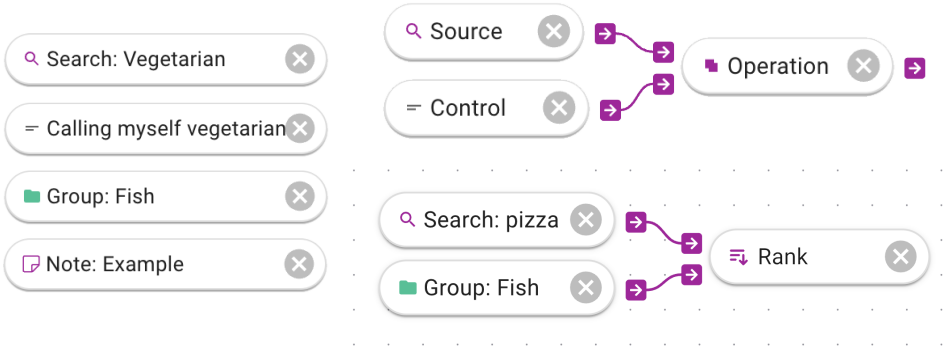


Figure 5.3: Sets of one or more documents are sources and operations are targets. On the left, sources types are shown, including single documents, keyword searchers, user-made groups, and arbitrary text documents called notes.

Workspace

We developed the Teleoscope workspace to reflect the process of arranging data on a table (**DG2**). The workspace is a drag-n-drop visual workflow editor which allows users to create computational graphs. We have two types of nodes in a workflow graph: document *sources* and *target* operations. We chose a direct manipulation metaphor for positive transfer from whiteboard apps that qualitative researchers are familiar with (such as Miro) and to mimic the process of arranging and rearranging documents on a physical desktop such as in affinity diagramming (see Figure 5.2).

Sources

A *document* is the primary interaction object in Teleoscope. Documents display a title, text and metadata information within the node or in the quick viewer when selected. Best performance for documents is 1-3 paragraphs, such as social media posts. Users can review documents to decide whether they are topically relevant by grouping, bookmarking, or using them as input to operations. A *search* is a fuzzy keyword search across the full document set; queries are in plain text and support set operations like *and*, *or*, and *not*. Keyword search was provided since this is a common search interaction style that qualitative researchers have often been explicitly trained in and are used to thinking about. Documents can be sorted into labelled *groups*, either by drag-n-drop on a group, or by selecting the group in a drop down.

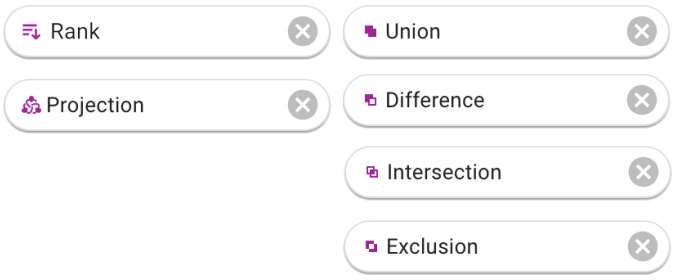


Figure 5.4: Targets are any operation on source documents. The left Rank and Projection operations use vector similarity to organize documents: into an ordered list for the Rank operation, and into clusters for the Projection operation. On the right, four set operations non-destructively join document sets. Set operations can also be sources.

A *note* can contain arbitrary text, which is then vectorized live as the user types. Users can use notes to create annotations, or they can use the note to drive Rank operations (see Figure 5.4).

Targets

Targets are the ML operations placed at the end of a workflow and produce collections of document collections as outputs. Each *source* will create a new document collection, whereas each *control* will manipulate distance metrics and similarity scores. Targets are meant to reflect the user's mental model as they explore the data so that the user can update their mental model along with the machine model.

The *Rank* operation ranks source documents relative to the average vector of all control documents (see Fig. 5.6). If no sources are connected to a *Rank*, the operation will rank all documents in the corpus; otherwise it will rank each source subset independently.

The *Projection* operation runs semi-supervised dimensionality reduction and *clustering*.

Control inputs to a projection define a distance metric: if two documents are in the same group, their distance is set to a minimum; otherwise, their distance is the cosine distance between their vector encodings. Users can choose to include an additional rule in the distance metric that sets the distance between groups to a maximum. Doing so forces the system to separate groups when clustering; otherwise, groups may be clustered together.

In the projection operation, *source* inputs can be used to define the

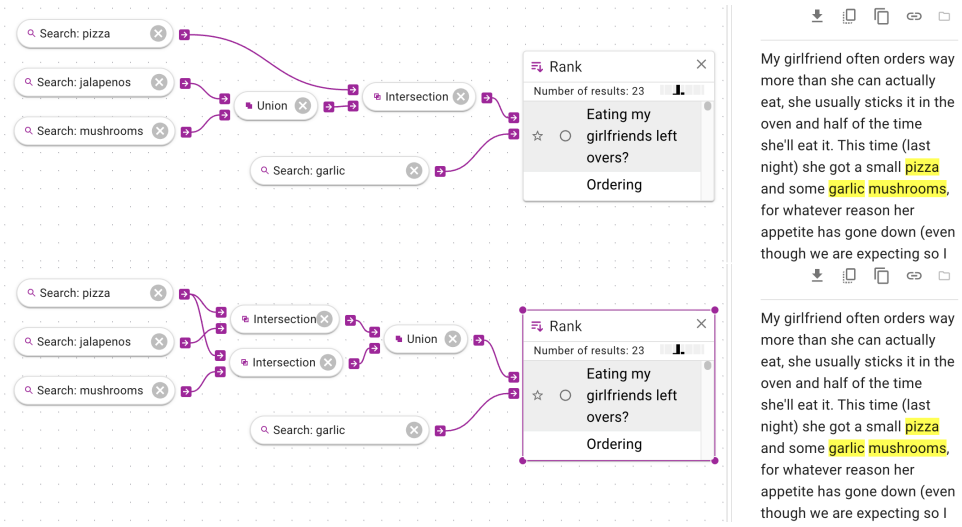


Figure 5.5: Long chains of set operations can be combined to drill down on document queries. Here is an example of two equivalent chains of set operations.

domain of the documents being clustered. If no *source* inputs are provided, users can decide between a random subset of the corpus, or a selection of documents ranked relative to the average vector of *control* inputs.

Source and target operation: Set Operations

Set operations are provided so that users may non-destructively combine sources and inspect results. They include standard set operations of union, difference, intersection and exclusion (the complement of intersection). They are the only operations which are both sources and targets; they can be combined to create chains of operations (see Figure 5.5).

5.3.2 Collaborative Curation Process

Users start with a dataset inside a workspace. They can create multiple workflows to explore and subset the original dataset. All workspaces can have multiple simultaneous users; all system state including window positions, node and edge connections, and Note content are synchronized with a central server and made available in real time to users in different locations.

A standard curation process would start with a keyword search. Users look through documents in a keyword seach and bookmark and/or drag

documents onto the workflow to indicate interest. A user can arrange these staged documents creatively, e.g., putting documents that are thematically similar together, making initial groups and dropping documents into them, or leaving documents unorganized until a later time.

The process of organization is intentionally messy. We wanted the first part of the interface experience to be a bit like affinity diagramming, where spontaneous document similarities can emerge without a computational or analytical commitment. At this point, users can show these proposed arrangements of documents or groups to collaborators to get feedback. They can take notes directly on the interface with the Note operation, and even use Notes to drive further exploration.

Part of the exploration process includes using the Rank operation. By connecting different example documents to a Rank, users can find more examples to add to their curated document groups (see Figure 5.9 to see an example of how a research team formalized their explorations). If a researcher finds that many documents from a Rank output are thematically different, they might add more examples to refine their theme. If the Rank has many thematically consistent documents, they might commit to their theme by copying the documents from Rank to a static named Group.

Projections are used when enough themes have been developed as Groups to create stable outputs from the clustering algorithm. There is no strict minimum number of Groups that can be used, but clustering algorithm outputs vary greatly for low number of inputs. High variation can be both an advantage and disadvantage, providing opportunities for discovery, but making it difficult to make strong claims about the corpus.

5.3.3 Common Workflow Patterns

Analyzing our own use of Teleoscope along with user data, we have come up with example workflows that demonstrate different document search strategies. These illustrate the creative potential for data curation using Teleoscope workflows.

Order Search by Source

Sources do not have an intrinsic order, however, Ranks can order a source by any other source. For example, a researcher might want to see which documents in a keyword Search are most like an example document. Or, they might want to order a Group by similarity to another Group. Ranking allows for different perspectives to be imposed on a source which enables

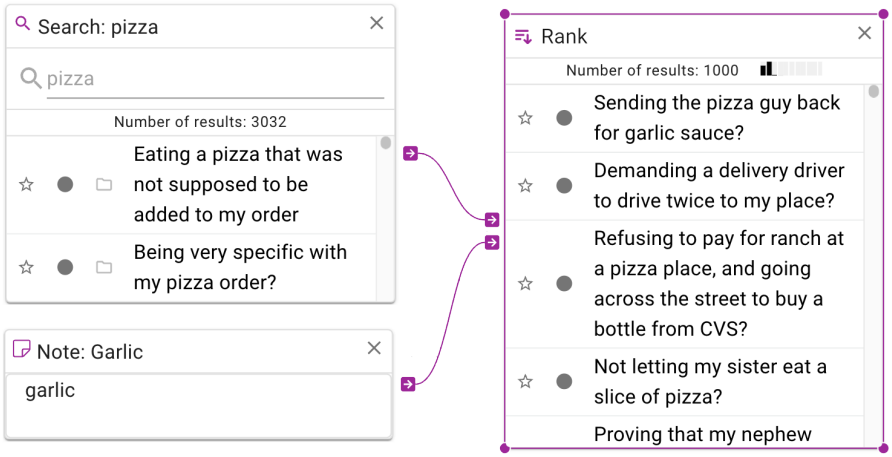


Figure 5.6: Ranks can be used to order any source. Above, documents in the search that are closest to the vector representing "garlic" show up at the top of the Rank.

a closer check to see whether themes are being accurately captured (see Figure 5.6).

Rank-Group as theme

The simplest pattern is to treat a combination of a Rank and a Group as a refined theme. As documents are added to the Group, the Rank updates and expresses a potential underlying theme. Documents from the Rank can then be added into the Group to further refine the theme.

Search-Difference to discover keywords

A single Group can encompass a large number of documents. To discover more keywords, the union of different Searches can be combined with a Difference operation to discover new keywords (see Figure 5.7). This workflow pattern allows for large groups to be closely analyzed.

5.4 System Architecture

Reflecting our goal of real-time interaction (**DG3**), Teleoscope's backend is engineered to support continuous iteration and interaction. This was a non-

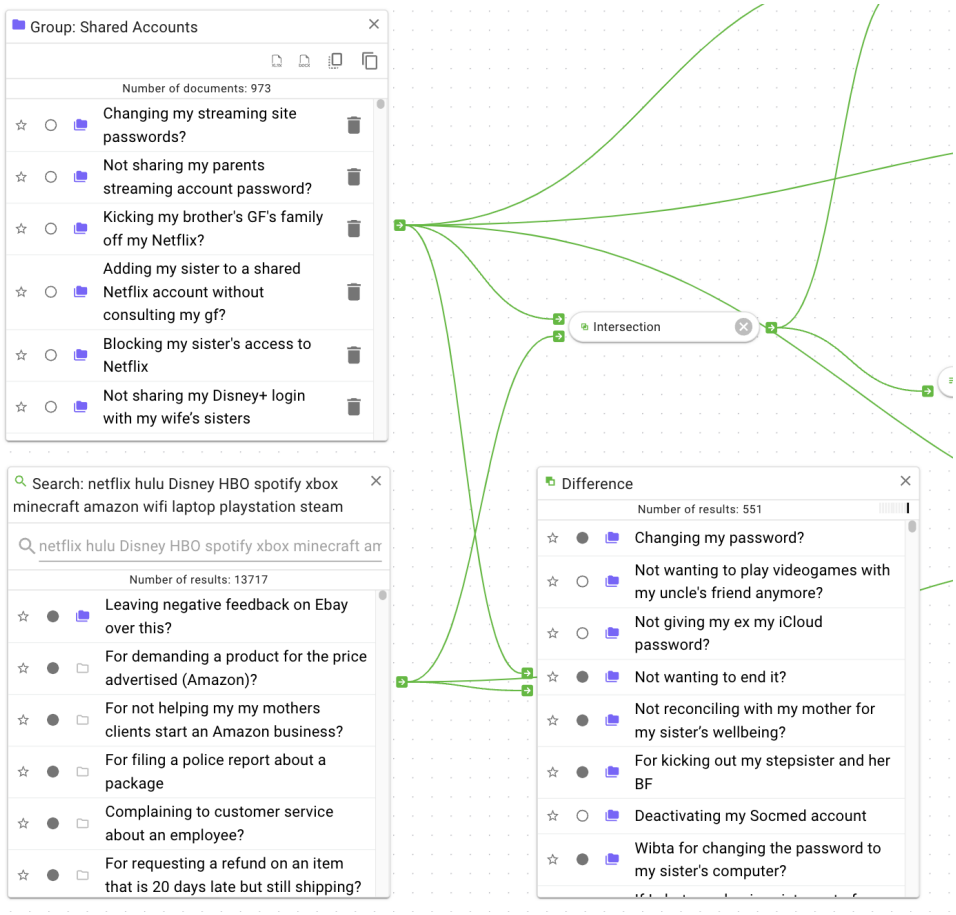


Figure 5.7: An example of using a union of keyword searches and a difference operation to ensure that all interesting keywords in a large group have been found.

trivial task, which is not always supported by other systems, due to the high computational workload. Through internal testing, we found a distributed backend was necessary to run Teleoscope computations without blocking user interaction. We chose to precompute, cache, and distribute as much as possible to make on-the-fly calculations seem quick. For example, a Rank of cached data is nearly instantaneous despite sorting hundreds of thousands of documents (average 1.6s for update to UI, including network latency). For longer-running operations, the interface allows continuous interaction with other elements. We believe that this user experience goal is vitally important to Teleoscope being capable of facilitating a creative curation.

From a community perspective, we wanted Teleoscope to be available for qualitative researchers who are non-computer experts, which necessitated a robust enough system to survive a production-level environment on the open internet, including almost-one-click deployment, user accounts, security, and backup systems. Teleoscope is continuously available live at [anonymized for review] for use by the general public. We have an active user community on Discord where we take bug reports and design requests.

Frontend

Our frontend is built using the NextJS ecosystem to manage React development and deployment, ReactFlow for the graph drag-n-drop workflow system, and Material UI for the design elements, as well as a variety of smaller libraries and custom components. NextJS was chosen because of its large user community and full stack support, including data fetching and user authorization libraries. We chose ReactFlow after experimenting with a number of whiteboarding, windowing, and drag-n-drop libraries; it is also a mature and actively maintained freemium product. Material UI implements Google's material design in React.

Backend

Teleoscope uses a distributed backend with RabbitMQ for messaging and Celery to execute tasks. We use Milvus for a vector embedding database and MongoDB for all other data. To ensure continual service, our system is daemonized with linux *systemctl* as well as the *pm2* library for node and python applications. Both perform process monitoring and memory management.

Dataflow and History management

React uses a virtual document - object model to ensure a strict dataflow model for user actions and system state. Similarly, we designed a dataflow policy such that the frontend (almost) entirely makes requests to the backend to manage system state on the server. This means that actions that mutate database state are strictly sent via a secure websocket connection to RabbitMQ through a well-defined API. With the exception of user registration, there is no direct database mutation by the frontend. Similarly, backend state updates are managed by the Stale-While-Revalidate data caching and fetching system in NextJS.

Keeping this strict policy has benefits and drawbacks, mostly having to do with interaction availability. Any large-scale calculations and mutations are offloaded to the backend while the frontend waits for data to be marked as stale to refresh the local client cache. The trade-off is that some state changes that require a backend response may be impacted by network latency.

5.5 Deployment Case Studies

This section outlines our evaluation processes. We evaluated Teleoscope through (1) informal piloting and internal analysis using low-cost evaluation methods; (2) a multi-week study and focus group with qualitative researchers (N=5) including a post-hoc expert review of our interface and study data with a visualization group; (3) a multi-month field deployment with a qualitative research group; and (4) an on-going public release. We used data from Reddit as archived by PushShift (Baumgartner et al. [2020]) up to their latest published data in January 2023. Research was conducted under approval of our institution's research ethics board; participants signed consent forms and were reminded of their right to halt participation in the studies if desired.

We focused on one subreddit, the *r/AmItheAsshole* advice forum since we were ourselves interested in data on social norms; our field deployment collaborators also wanted access to *r/nursing* and eventually their own arbitrary data uploads. *r/AmItheAsshole* has roughly 650K documents and *r/nursing* has roughly 100K documents. Data included only posts, not comments. The last post date was February 2023, since Reddit significantly restricted their data API access in response to ChatGPT.

5.5.1 Case Study 1: Piloting

In the first year, Teleoscope first went through a series of very quick design iterations within our team. As we developed it into a more robust and large-scale system, we began to incorporate more human-centred design techniques into our design process as our research and development team grew. We are including only light details on our informal methods so as to faithfully report on our process. During this year-long phase, we also ran a series of tests on different NLP approaches. These were incorporated into our system during our informal user testing.

Participants

For our informal methods, the users we refer to are members of our design team and the larger lab members who were not involved in Teleoscope development. In terms of expertise, all users were trained competent-to-expert computer scientists; some members of our team are trained competent-to-expert UX and qualitative research practitioners.

Methods

We used a variety of informal low-cost UX evaluation methods to motivate our early design choices, including cognitive walkthroughs, heuristic evaluations, and informal observations with both people from our design team and from our larger research lab. The cognitive walkthroughs and heuristic evaluations were performed with standard methods with heuristics taken from the Nielsen Norman group (Nielsen and Molich [1990], Nielsen [1992], Neilson). Informal observations were performed on low-level interactions such as menu clicking and basic keyword searches to discover and amend heuristic violations.

Results

The results of our initial evaluations were a set of guiding backend and frontend design requirements that aligned with standard UX heuristics. We summarize the most relevant heuristics here to explain our early design directions:

Visibility of System Status. Our original design used a dashboard metaphor where each module displayed system state such as included/excluded keywords, document similarity statistics, and topics. Our initial corpus visualization attempts repeatedly pointed towards common visualization solu-

tions of weighted adjacency matrices of keywords and documents, but with corpus sizes of a million documents, pixel overlap became a problem very quickly. Further, the connection between modules in a dashboard is hidden. Therefore, we decided to move towards a windowing system, eventually creating window modules that had visible input/output areas.

Recognition over recall. By moving commands and system state out of menus/collapsible dashboard modules, we opted for a design with minimum display of information, contrary to our prior approach that displayed summary statistics of keyword distributions. We deemed this unnecessary for our core interaction goal; opting instead for an Overview/Details-on-Demand design pattern where the *process* was visualized rather than the full system state.

Error recovery. We created a robust history system where every system action is logged. After many discussions about how much history to display to the user, we opted for an algebraic workflow metaphor. This way, the user can directly manipulate the "history" of their actions.

Case Study 1: Conclusion

During this study, we quickly attempted and discarded visualization approaches that are common in topic modelling, including displaying and interacting most directly with keywords and topic labels. Instead, we moved towards a direct manipulation approach, where documents and outputs were visualized as arrangable windows on a workspace.

5.5.2 Case Study 2: User study and Focus Group

Once our front and backend designs had stabilized, we released our first version for a multi-week study with representative target users. The premise was to simulate a research team working on the same research question within a provided dataset. We were interested in the following research questions:

- **CS2.RQ1**. In what ways did participants understand and use features such as collaboration, search and ranking?
- **CS2.RQ2**. How did participants incorporate Teleoscope into their understanding of qualitative research processes?

We were also interested to see the extent to which Teleoscope could hold up in a simulated production environment and welcomed ongoing bug

reports. Therefore, the interface was developed to be robust enough for participants to use on their own devices, outside of a lab environment.

Participants

We recruited participants who had competence with qualitative methods: at least an upper-level undergraduate course and/or equivalent research experience. Eight participants were recruited; three dropped out (N=5 final count). Of participants who remained, one was a PI at a university who leads qualitative research in Nursing, one was a senior PhD student in Sociology, and two were upper-level undergraduates in Sociology attending a directed studies in qualitative methods course, and one was an upper-level undergraduate in Psychology. Participants were reimbursed at a rate of twice local minimum wage due to their status as expert users.

Methods

Participants were introduced to Teleoscope and each other during a one-hour training session where we brainstormed a shared research topic. Then, participants were instructed to use Teleoscope for at least 10 hours before the focus group, scheduled three weeks later. For each session that they used Teleoscope, they wrote in a diary, detailing (1) the *theme* that they explored; (2) the *process* by which they explored the theme; and (3) any *collaboration* features they used; and (4) bugs or features requests. No design changes were made during these weeks, but minor bugs were fixed. System logs were kept throughout this period.

Diary entries were analyzed using affinity diagramming before a focus group with participants, which took three hours including lunch. Focus group involved: (1) diary discussion; (2) brainstorm on problems encountered; (3) brainstorm on feature requests and design solutions to problems; and (4) explanation of ML concepts and a brainstorm on better alignment between visual and interaction metaphors. We video and audio recorded the focus group and used large printouts of the Teleoscope interface to draw and annotate problems and design ideas (see Fig 5.8).

Post-hoc Analysis by Visualization Group

After we had analyzed and summarized participant results, we presented Teleoscope in two multi-hour analysis sessions with a Visualization research group. Our results reflect the analysis of that group along with our own analysis and solution brainstorming.

Figure 5.8: Large printouts of the Teleoscope interface were used to draw and annotate problems and design ideas with both the focus group (left), and the visualization group during a post-hoc analysis (right).

Diary Results

The topic that was chosen by the group of study participants for investigation was *Critical and end of life care* within the *r/AmItheAsshole* dataset. The group brainstormed starter keyword search ideas of *Medical Assistance in Dying (MAID)*, *End of life care*, *Palliative*, *ICU*, *Failures*, *Emergency rooms*, *Emergency care*, *Lack of beds*, *Overcrowding*.

Teleoscoping differs from keyword searches. Participants found that the ranking system differed from a normal keyword search. For many participants, it took some time to (1) differentiate results of a keyword search from results of a Rank operation; and (2) differentiate valid results that did not meet their expectations due to the documents that existed in the corpus from invalid results due to bugs or mental model inconsistencies. There were negative transfer effects from being used to keyword searches which took multiple sessions to unlearn.

For example, P1 searched for "palliative" and was "*...surprised by how many posts were about animals at end of life, which does not fit our defined research topic.*" P1 then wished "*...there was a feature that would take everything I had already put within one group and give me 'more like this'*", which was exactly what the Rank operation was designed to do. Multiple participants corroborated this sentiment in their diaries (P5, P7, P8). This indicated a problem in the participants' mental model, likely due to (1) how we were representing the Rank operation on the workspace; (2) our training, documentation and support materials; and (3) not enough time to learn the tool.

However, P1 reported for their third session that they spent a long time looking through documentation and support videos to understand the possibilities of Teleoscope:

> *Today I also spent time trying to go down the rabbit hole of different searches to try to gain a true appreciation for how this machine learning approach to data collection differs from just keyword searching within the Reddit search. This was really evident to me when I found a post where the OP had a palliative/terminal illness, and I wanted to find others where this was the case. I made a new folder for this category, then used the [Rank] feature, and immediately found one other post where the OP has cancer and was asking a friend to not mourn their death. It would be extremely difficult to keyword search for this type of topic, but it's a very interesting and important area to capture (OPs with terminal illnesses). This was a great exploration! (P1)*

This indicated that it was possible to learn the difference between keyword search and Rank, but that the learning curve was steep enough to require multiple hours of usage and documentation review.

Teleoscope can support quick, iterative workflows. P7 articulated a very clear document review strategy and seemed to understand the tool very quickly. Ignoring the group topic, they searched for documents related to their own research program by skimming titles:

> *I was interested in [AITA] posts about gay marriage, which is a topic tangentially related to my own research. I populated the [gay wedding] group with results whose titles caught my eye. I should note that I very rarely read the actual documents. If the title was vague, I occasionally skimmed the first few lines. (P7)*

They further suggested adding a document quick viewer to aid in skimming. They then organized documents into groups, relabelling and changing the groups as they developed their understanding of the corpus and the tool. Then, they switched between using the Rank operation and the group feature to find relevant documents:

> *After adding about half a dozed documents to the group 'gay stuff', I noticed that many of the documents are about gay panic. That is, the fear of being (wrongly or correctly) as gay, the dislike of anything perceived as gay, and an aversion from being around gay people. I changed the group name to 'gay panic' to reflect this...once I had about 13–14 results, I opened the [Rank] window for the group. Looking at the first two pages of results, none of them that weren't already in the group seemed very relevant, mostly judging by their titles and occasionally by the first couple of sentences in the document. I refined the search...[with a] couple of documents that I thought particularly demonstrated gay panic. (P7)*

Positive and negative transfer effects from other qualitative research software. Participants' prior extensive experience with qualitative research software allowed them to have a much more clear mental model of the tool without extensive training, which indicates the possibility of positive transfer effects. Many desired features were given as examples from tools that they had experience with, such as Google docs, MaxQDA, and NVivo. Unsurprisingly, they mostly expected Teleoscope to work like other

interfaces they had previously used. To our surprise, none of the participants used any of the collaboration features.

P7 noticed many problems with the interface and made many suggestions for design changes that we brought to the focus group, including a lack of annotation and coding features, document export features, and overall corpus visualization features: "*As this point of my exploration of this theme, being able to play around with how [documents] connect to one another would I believe might have helped refine my thinking.*"

The above results were summarized and presented at the focus group, motivating our central discussion points.

System Log Results

During the user study, the system maintained a log of user actions as they interacted with the system. Across the 5 participants, 7 sessions with the Rank operations were tracked (one participant created 3 separate sessions for themselves). The mean number of actions tracked per session was approximately 310 (median = 286, minimum = 97, maximum = 684). Actions included such things as session initialization, creation/movement/deletion of windows, keyword searches, instantiation of ranking and results, in short, any conditions where the state of the user workspace was altered.

The actions were logged and then visualized to better understand user interactions. Generally, users made use of an iterative process to find documents of interest, alternating between putting documents into groups and instantiating new Rank operations to find new documents relevant to their queries and then sorting them into their groups. The number of Rank operations created across the study were relatively small however (between 2 and 4).

Focus Group Results

The focus group provided insight into the needs of qualitative researchers with different levels of expertise with computer supported analysis. We used thematic analysis to review the results. We report here on the most prominent needs that emerged.

Mental model need: Rank state needs to be inspectable. In our tested design, Rank state was coupled directly to a single group. When the group was updated, the Rank changed. Individual documents within the group could be selected to weight the Rank search vector closer to that document. This confused participants about the state of the Rank. We

decided that the Rank inputs and outputs needed a more explicit visual representation and decoupled from a single group, allowing for multiple input sources.

Mental model need: Use direct manipulation for all features, including workspace interactions. Participants expected more features to use direct manipulation such as drag-n-drop, infinite canvas, and organizing documents.

Mental model need: Clarify the exploration metaphors onscreen. For example, participants wondered whether they were being "*dropped off in the landscape*" of documents and "*going down different paths*" (P7). If so, they wanted a record of the paths and some way to compare paths directly onscreen. Participants agreed that "*seeing it*" and "*understanding how close documents are to each other in space [is important].*" (P1)

Feature need: Support set operations and filtering. Participants were most familiar and received direct training in keyword manipulation. Therefore, they were familiar with set operators and wanted to use them to gain confidence that they were being thorough enough in their search.

Conceptual need: Confidence in path saturation. Participants agreed that they did not need to show total path *exhaustion*, rather, they needed a sense of *saturation* and demonstrable information power. Their imagination of the use case for Teleoscope was for data curation, which did not mean finding every piece of relevant data but instead finding enough *representative* samples of data.

Conceptual need: Explaining the methodology to reviewers. Participants were concerned about how to explain the Teleoscope process to reviewers at a high level, but were not concerned with the details of the Teleoscope process. As long as they had a clear metric that they could point reviewers to (e.g., a paper that describes the metric), they did not particularly care about which metric to use. This was a defensive publication strategy: in some qualitative papers, keyword searches are reported on directly to demonstrate that they had reached saturation or sufficient information power. Large data sets and less of a value of statistical rigour for reviewers means that any reasonable metric could be argued for and used.

Conceptual need: differentiating data curation from analysis. Participants were unsure whether using Teleoscope constituted a violation of rigour such that it mixed data curation and analysis research phases. Particularly if text-level analysis was to be supported, they felt that it might not be appropriate to allow for both in the interface at once. We found that participants used Teleoscope as we intended: they explored data in a manner that extended beyond keywords searches. For example, participants

reported the following:

> *I was excited when the tool came up with things about the topic, but not including keywords that I used. (P1)*

> *I use it to find papers that I wouldn't have found. (P7)*

This allowed us to believe that Teleoscope could be used for a longer and more in-depth research project.

Post-Hoc Analysis and Recommendations from Visualization Group

We presented our results and current interface to a visualization group at our university. After two multi-hour analysis sessions, the visualization group recommended the following:

Visualize workspace relations. Up until this point, our interface did not include any visual graph concepts since our original abandoning of adjacency matrices. The recommendation was to clarify system state further by treating windows as nodes which could be used as input sources for operations.

Clarify data type representation. Since the results of workspace operations were effectively variations on ordered lists, the recommendation was to create visual homogeneity where data types were the same, and visual differentiation where they were different. From this, we developed our current paradigm of source, control, and target data types, and a unified output datatype of a list of document lists.

Create a quick-viewer to allow for minimized windows. A sidebar was recommended to allow documents to remain as minimized pills while users reminded themselves of document contents.

Allow for multiple workflows. Users wanted to simultaneously pursue multiple explorations with quickly accessible workflows. They further recommended that we create chainable workflows to target both provenance and reusability.

Our main takeaways from the visualization group recommendations was to move our window metaphor into a workflow metaphor with simplified workspace objects. This required a major redesign for both our front and backend systems to support graph operations. It also introduced questions of concurrency and graph directionality. For example, our first imagination

of this design introduced cycles into the graph; as such, we made the decision to make sources strictly "left-handed" and targets "right handed" with the exception of set operations. This had further impacts on our state management system, which had to be redesigned to work with a graph-based structure.

Case Study 2: Conclusion

In this study, we investigated how real qualitative researchers would use Teleoscope in a simulated research environment. A summary of our findings corresponding to the research questions is:

- **CS2.RQ1 Result**. Participants understood the overall concept of the Ranking operation and workspace as a whole, likening it to walking down different paths in an unknown region. Even though this metaphor came from the participants, the operation inputs and system state were unclear; therefore we decided that the path metaphor should be visualized as directly as possible. Participants did not use any collaboration features of Teleoscope, which may be due to a lack of impetus, our study environment, or because of our design. In the next study, we addressed these possibilities to ensure collaboration features were used.
- **CS2.RQ2 Result**. Participants interpreted Teleoscope primarily as a data curation tool. Further, they were unsure about whether it infringes on analysis and had concerns about academic rigour if it does.

5.5.3 Case Study 3: Field Deployment

After completing a redesign after Case Study 2, we were interested in a long-term focused field deployment with qualitative research groups who could bring their own research needs to us. We also wanted to move Teleoscope out of a simulated environment and into a production environment where real-world motivations and difficulties would be encountered.

Case Study 2 had many study contrivances, such as a research topic that none of the participants were using for their own research. We felt that it was important to see how Teleoscope would perform when participants were subjected to all the benefits, consequences and costs of real research. We

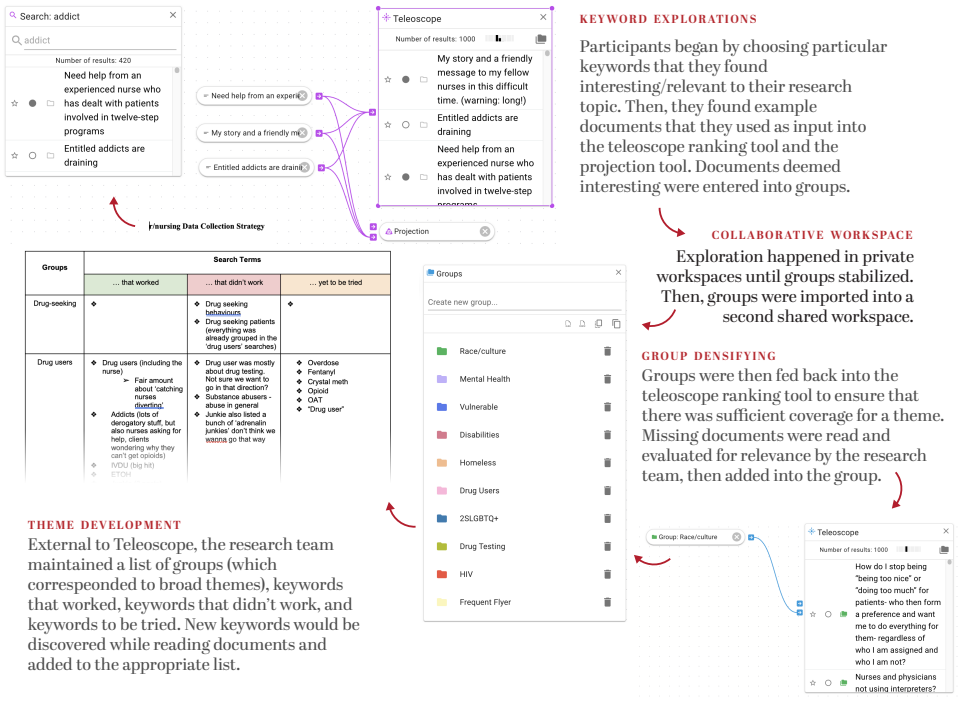


Figure 5.9: An example of a workflow from *Case Study 3* our long-term deployment. Pictures are actual screenshots of research artifacts from our participant research team's data-gathering phase. Participants worked both individually and collaboratively on the Teleoscope interface, and collaboratively on Google Docs, Zoom, and in-person. Due to the existing qualitative research culture in their research group, keyword searches were the focus of their data curation approach. ML features were used to discover new keywords, find thematically similar documents that did not have specific keywords, and to saturate groups with relevant documents.

also committed to ongoing feature development (and bug fixes) using a quasi co-design methodology.

Specifically, we were interested in the following research questions:

- **CS3.RQ1**. When put in a real-world environment, how did researchers incorporate Teleoscope into their own research practice? Which features, processes, and workflows within Teleoscope were commonly used and in what way?
- **CS3.RQ2**. Using Teleoscope, to what extent were researchers able to feel confident that they were able to retrieve data using criteria that are important to qualitative researchers, i.e., richness, information power, saturation, etc.?

During this period, we also moved Teleoscope out of a test environment into a publicly-available production release. This involved adding many security features and backup systems which were exposed to the threat of arbitrary internet attacks. We had at least one successful security breach which was dealt with and mitigated; the number of unsuccessful attacks are unknown.

Participants

Three PIs and research groups were recruited; only one research group was able to commit to long-term use of Teleoscope. The final team was comprised the PI from Case Study 2 and a research team of two graduate RAs that were recruited specifically to use Teleoscope. Participants were not reimbursed for their time by us since they were being supported extensively by our design team for their own research project; our understanding is that RAs were compensated as normal by the PI.

Methods

Teleoscope was deployed in a standard beta release manner where participants were given a private link to Teleoscope until we transitioned to a full public release. Participants were invited to participate in a Discord server to make bug reports and feature requests. Depending on the request, occasional emails and video interviews were conducted. Logs were kept of system use to compare with study results. For the purposes of this paper, we finished data collection after six months, but use is ongoing by the research team.

Results

The research team used Teleoscope for data curation for their research project on nurses and structural inequality as articulated in Reddit's *r/nursing* forum, which is where working nurses post about their day-to-day problems. Of their own initiative and in alignment with their own qualitative methodological approach, the team's main conceptualization of data curation was via an external google document that contained a set of keywords to be used in searches (see Figure 5.9). After exploring the data using a variety of features in the Teleoscope workspace (described below), they would update their keyword list with successful and unsuccessful searches. Here are the ways in which the research team used the Teleoscope workspace features:

Teleoscope helped discover unknown terminology. Even though the research team was composed of people experienced in nursing culture, they were not always aware of the terms that were used by on-the-ground nurses.

> *We search based on these weird, predefined words keywords that we think relate to structural inequities. But we also have to guess in advance what language people might be using... We didn't think that people would delicately [post using the term] 'people who experience structural inequities'...But we did try a lot of words we wouldn't use, you know, like addict, junkie... We were trying to use the system in a way that would get us further than those keywords alone. (P1)*

The Teleoscope system helped to populate their keyword search document with search terms that they would not have predicted a priori.

Teleoscope helped with search saturation. The researchers reported (1) making groups from documents from their keyword searches; (2) piping the groups into the Ranking operation as controls; (3) determining which documents they had not yet read and (4) adding those documents to the groups. This helped to see the parts of the document space that they had not yet captured with a keyword search.

> *Putting each one of our single groups like indigenous, vulnerable disabilities, [an RA] put them into the [Rank] and then basically went through to see like which ones we hadn't read... We were just trying to expand our data set and be exhaustive. (P1)*

Working iteratively between keyword searches and ML functionality was important for exploring and structuring the research topic. The researchers reported that they used the system to iterate and structure their ideas about their research topic.

> *It was actually really helpful to start with keyword searches. And to be able to build out this groups structure, and then [Rank] from there. Whereas maybe if we'd had a more like drilled in topic, we could have just gone from there.(P1)*

The researchers used Teleoscope to develop ad hoc themes out of their original categorical approach as their understandings of the target data grew:

> *There are a bunch of different like origin categories we needed to go off of because of the way our topic is. We couldn't even [try to search for categorical terms such as] 'Oh, this is about emergency departments...vulnerable [people] or inequities'...because when people are dragging on someone who uses drugs, who comes to the emergency department every week. That's not the words they're going to use. They're going to use super stigmatizing language, probably like what we've seen in a lot of cases and be like, 'I'm so frustrated. This junkie comes into work all the time. He's just drug seeking. He's like plugging up a bed for everyone else who needs it.' (P1)*

Projections have potential to allay methodological concerns. In the focus group, it was brought up as a concern that using Teleoscope might be too close to analysis and confuse methodological rigour (in rigorous qualitative research, data curation and analysis stages are kept distinct so as not to predetermine results). However, the projection operation, which was added after the focus group, seemed to have potential to populate the interface with unexpected results:

> *One of the things that I had really worried about methodologically was that with Teleoscope, I almost felt like you were kind of like deciding what your findings might be...I feel like [the projection operation] is really addressing some of that for me, because I feel like it's bringing you all these like adjacent topics to what you're looking for. And I feel like it really broadens your idea so much further in the potential data [since] you're exposed to so much*

> *more of the subreddit than you would be through just keyword searching. And I feel like that's really methodologically sound. (P1)*

The speed and ease of use of the Projetion operation helped with researchers' sense of completion as well:

> *When it comes to big data sets, it takes way too long to get a sense of what's going on. [The projection operation] is really nice, quick [and has] potential to expose that kind of stuff so that your findings aren't [close to the] single keyword that you put out...With qualitative research that's a really interesting and powerful thing to be able to do. (P1)*

Multiple workflows/workspaces allowed for reproducible exploration and collaboration. The research team approached collaboration by exploring on their own in independent workspaces, discussing their results, and then collating the results into a single shared workspace. Outside collaboration tools were also used, such as email, Zoom, and Google Docs. The independent workspaces served as drafting areas, where individual researchers could explore many ideas without committing to the larger team's conception of the dataset, then came together with refined groups and keywords. After they collated their results, they performed further data exploration as a team on a single workspace by systematically using our ML operations to ensure completeness.

One of Teleoscope's key design features is to create reproducible workflows that are able to be inspected by collaborators. The above method of exploring in separate workspaces and combining in a shared workspace was enabled by the guarantee of maintaining reproducible results.

Along with the above findings, researchers submitted a variety feature requests and bug reports. Set operations were the last feature to be developed after some discussion and redesign of our backend graph processing system.

Case Study 3: Conclusion

In this study, we performed a customized field deployment for a real qualitative research team and took ongoing bug reports and design feedback. The summary of our findings for our research questions are:

- **CS3.RQ1 Result**. Researchers incorporated Teleoscope into their research practice as a data curation tool, working between an external Google doc for keywords and the interface itself. Teleoscope was used to explore parts of the document space where keywords would not be easy or obvious to find. The most common features used were the keyword search, document reading and grouping. [Ranks] were used after grouping. Projections were used after Teleoscoping and grouping.

- **CS3.RQ2 Result**. Teleoscope helped to provide confidence that a corpus was being more fully and rigorously explored by providing both ranked and randomized example documents.

5.5.4 Case Study 4: On-going Public Release

We are hosting an ongoing public beta release of Teleoscope (publicly deployed at [anonymized for review]). This is not a formal user study; instead it is an ongoing test and demonstration of our system robustness in terms of performance, security, and availability. By committing to a live release on a cloud platform, we were forced to develop the following security and availability measures:

- **(Nearly) one-click deployment**. Teleoscope can be deployed on a new AWS virtual machine using Ansible playbooks in nearly one click. This was developed after a sprinkler accident destroyed our original non-public servers and motivated our move to the cloud. We needed to re-deploy Teleoscope often enough to spend time developing an automatic deployment system.

- **Robust backup system**. We expected a catastrophic security breach at some point and developed a backup system. When our database was indeed hacked and erased, we redoubled our backup system to two small-scale hourly offsite backups as well as a daily backup.

- **User roles/API limiting/Reverse proxy/SSL/TLS**. Earlier this year, Reddit restricted usage of the data API, which interfered with our data collection strategy. It also introduced the threat of large-scale data scraping from Teleoscope. As such, we restricted Teleoscope to registered users (open to anyone to register), creating a robust internal/external user role scheme for MongoDB and RabbitMQ, put in place data security measures such as limiting our API throughput, and set up a SSL/TLS reverse proxy to encrypt messages between client and server.

Teleoscope remains online as of the writing of this paper and continues to gain users.

5.6 Related Work

In this section, we discuss the similarity and differences between Teleoscope and other related works. With Teleoscope, we departed from standard topic modelling because we are not interested in creating *categorical* topics, but instead are interested in supporting researchers in the first stage of developing rich, specific *themes*. The difference between categories and themes is subtle, but it is important to qualitative researchers. Rather than creating deductive categories, qualitative researchers are interested in inductive interpretation.

The difference is illustrated well with *Cody*, Rietz et al.'s excellent qualitative coding support tool (Rietz and Maedche [2021]). *Cody* addresses the problem of annotating specific sentences with qualitative codes (similar to tagging) by connecting keywords with set operations to create inclusion/exclusion rules. These rules are then applied across a corpus. This supports a style of coding that is fundamentally categorical, that is, the main action of analysis is to organize the different sentences into code categories. This is helpful in inductive coding, and is seen in new LLM-supported tools for qualitative analysis and sense-making (Gao et al. [2024], Kim et al. [2024], Liu et al. [2023]). Telescope, by contrast, is *not attempting to support users in coding and labeling documents*; instead we focus on curating document sets by visualizing provenance relationships.

On the categorical side, many topic modelling visualizations take the approach of acting directly on topic as document categories. For example, *Serendip* (Alexander et al. [2014]) and *TopicSifter* (Kim et al. [2019]) both approach the problem of topic exploration and discovery by allowing users to filter and/or drill down into topics. For these tools, the topics are the primary objects of interaction, and the goal is to explore the topics themselves. Both are a top-down approach. Teleoscope is most conceptually similar to *Scholastic* by Hong et al. (Hong et al. [2022]) because we also take a human-centred approach and provide a hierarchical clustering algorithm as a "machine-in-the-loop" approach. Their elegant approach of visualizing the document space differs greatly from our choice of visualizing the process of discovery. Also, as with *Cody*, they focus on document coding, which we do not.

Teleoscope differs from the above tools in that it does not attempt to

visualize, categorize, or make claims about the entire corpus. Instead, we take a bottom-up inductive approach which allows for orders of magnitude larger exploration, but with orders of magnitude less to display. In this way, although we use semi-supervised topic-modelling NLP methods (explained below), our conceptual approach does not manipulate topics directly; instead the user interacts with the ML model implicitly while interacting with documents and groups similar to how they would typically with qualitative analysis and thematic exploration. Put another way, we are making mini-topic models from of a restricted subset of documents based on customized distance metrics that we are calling *themes*.

Teleoscope differs from topic modelling tools by focusing on (1) discovering relevant documents rather than making claims about a whole corpus; (2) supporting collaboration, reproducibility, and schema expression; (3) co-developing themes through creative data exploration; and (4) allowing users to interact in real time with a larger corpus than other tools (**DG1**, **DG2**, **DG3**).

5.7 Discussion, Limitations, Future Work

In this section, we discuss the results of our studies with regards to our design goals, research questions, and provide directions for future work.

Through our studies and design work, we settled on a process-focused graph-based workflow metaphor rather than a dashboard metaphor (**DG2**). This was due to our focus on the process of arranging and connecting documents as being the most important interaction focus (**DG1**, **DG3**); it also increased visibility of system state and allowed for direct inspection of computational elements. Researchers found that they were able to find confidence that they had searched "enough" of the document space through using the Ranking operation, and found potential in the Projection clustering operation in supporting rigorous document space exploration (**CS2.RQ2**). After we iterated on our design to allow for collaboration and make it obvious to our participants how to use collaboration features, they were able to satisfy their need for theoretical saturation/information power through collaborative iterative searches using Teleoscope (**CS2.RQ1**, **DG2**).

A limitation of a workflow metaphor is that it does not visually model the *unknown* document space very well. In part, this helps manage the cognitive overload since users are only shown what they have explicitly searched for themselves. However, some dashboard elements such as summary statistics

of corpus exploration relative to selected groups would be an obvious next design step.

Due to our users' use of an external keyword Google doc, an open question is how and whether to incorporate keywords back into Teleoscope as a primary interaction element (**CS2.RQ1**). Perhaps with more experience with Teleoscope, set operations would be used to manage keywords more directly. However, we do not imagine Teleoscope as an end-to-end analysis tool; instead, we imagine Teleoscope as part of a qualitative research tool ecosystem. Teleoscope can work within existing analysis workflows between tools such as NVivo or PowerBI using our import/export feature to common document formats such as XLSX, DocX and JSON. Researchers were able to creatively include Teleoscope using external tools and to iterate between them and Teleoscope.

In our opinion, a large part of Teleoscope's success is demonstrated by its on-going public release and use by the research team that we recruited for field deployment in Case Study 3. We are continuing to recruit more research teams and hope to see Teleoscope develop into a more mature product as they use it.

Teleoscope's impact on methodology. One of the most interesting and relevant questions for Teleoscope is the way in which it can impact qualitative research methodologies. This was brought up by multiple participants, and is an important question for both adoption of Teleoscope and for shedding light on other burgeoning ML-supported methodologies that deal with data interpretation.

Current qualitative methodologies place a high level of importance on disciplined and ethical data curation, analysis, and reporting. Since interpretation necessarily comes from the perspective of the researcher, researchers attempt to maintain distance from analysis until data curation is concluded. Teleoscope could be seen as presupposing a thematic structure before detailed analysis has begun, which would violate the principle of letting the themes emerge from the data.

However, our participants articulated this mostly as fear of cognitive bias about the data, not that it fundamentally interfered with having an opinionated approach to data curation. It is impossible to do data curation from a truly unbiased perspective, since researchers still have editorial discretion in choosing their research interests, topics, and data sources. Our participants were satisfied with Teleoscope bringing up *unexpected* results along with *expected* results as a way to counteract bias.

We believe that Teleoscope provides an opportunity to directly inspect and compare biases by inspecting and comparing the workflows that trace

how a document collection came to be. Returning to the Nowell quote from Section 5.2, we believe that Teleoscope can demonstrate that data *collection* has been done in a precise, consistent, and exhaustive manner through *tracing the data curation process* with enough detail to enable the *research team* to determine whether the process is credible (**DG1**, **DG2**).

5.8 Conclusion

In this chapter, we presented Teleoscope, a web-based system that supports interactive exploration of large corpora (100K-1M) of short documents. We developed it in response to the need of qualitative researchers to explore large corpora in meaning-based ways using natural interaction techniques. Teleoscope provides ML-based workflows that have semantic and computational meaning. These workflows help researchers to retrace, share, and recompute their sense-making process. We reported on the design, engineering, evaluation, and deployment of our system. Our public deployment of Teleoscope is ongoing. We plan to continue improving Teleoscope and to maintain it for use by the broad community of qualitative researchers.

Chapter 6

Schema Crystallization

This chapter presents our analysis of the methodological process of using Teleoscope. The original paper was presented in a pictorial format; we have attempted to recreate that format here.

6.1 Overview

Large language models (LLMs) have very quickly enabled semantic text processing of large corpora in the range of thousands to millions of documents. The question of how to incorporate machines into a collaborative qualitative analysis with large corpora is still open. Collaborating with other researchers and the machine is difficult partly due to the black-box nature of the LLM. Yet, as qualitative researchers, having a healthy understanding of how an interpretation arose is as important as the interpretation itself. In this pictorial, we present a methodological analysis of collaborative LLM-assisted large corpora thematic exploration by tracking our use of a previously-developed LLM-based tool called Teleoscope. Teleoscope helps researchers externalize their internal cognitive schemas while making sense of data by capturing the process of interpretation in visual workflows. We contribute schema crystallization, a new concept that helps to integrate LLMs into a rigorous qualitative methodology for analysis of large corpora.

6.2 Introduction

Large language models (LLMs) have very quickly enabled semantic text processing of large corpora in the range of thousands to millions of documents (Asudani et al. [2023], Dai et al. [2023]). As a result, machine learning (ML) supported qualitative research is burgeoning; but, along with the promise, come serious methodological questions. Previous work shows that qualitative researchers are cautiously approaching ML support in thematic exploration (Bucci et al. [2024], Gao et al. [2024, 2023], Jiang et al. [2021]). The culture of qualitative research focuses on theory, positionality,

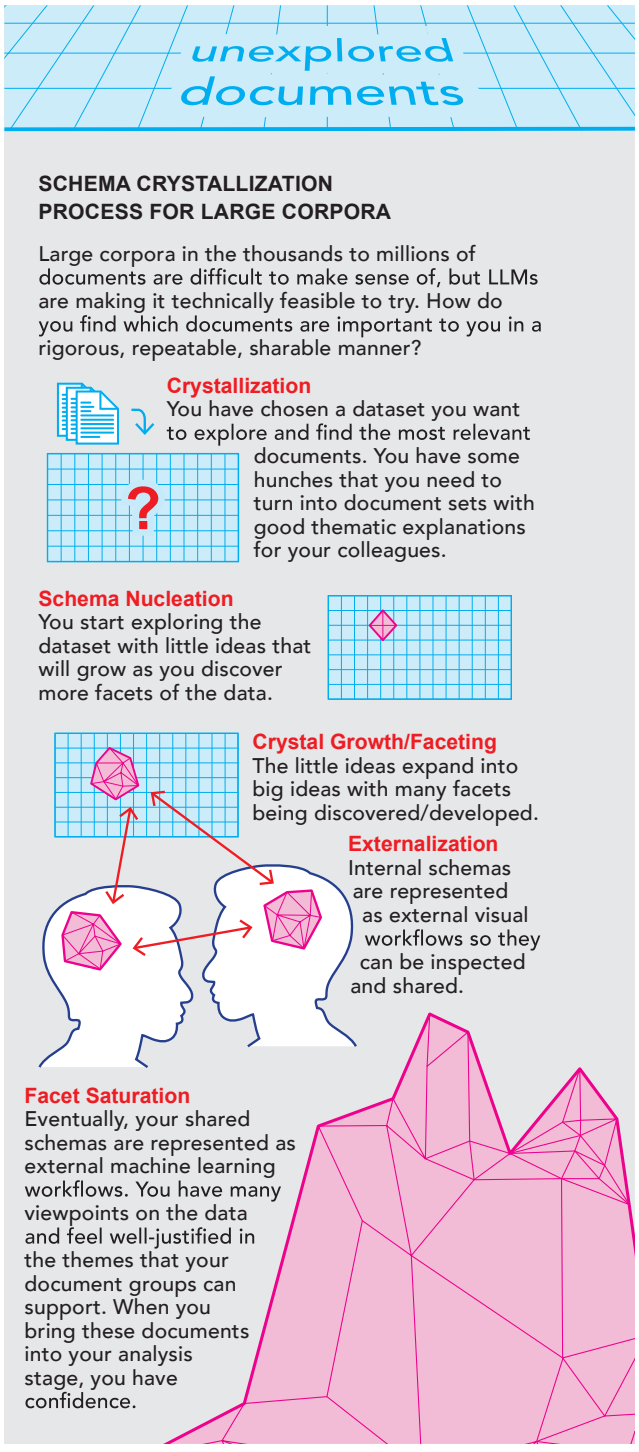


Figure 6.1: Schema crystallization introduction graphic.

and methodology to maintain rigour and protect against approaches that lack explanatory power and interpretation, or make unsupported conclusions (Hennink et al. [2017], LaDonna et al. [2021]). How to incorporate machines into a collaborative qualitative analysis of large corpora remains an open question.

In this pictorial, we present a process that we call schema crystallization (see Figure 6.1). This process developed while analyzing a large corpus using Teleoscope, a previously-developed ML-based thematic exploration tool (Bucci et al. [2024]). Thematic exploration is meant to be an analogue to the well-known qualitative process of thematic analysis but applied to the editorial process of exploring a dataset (Braun and Clarke [2022]). Schemas are the "way we think" about a dataset (Berret and Munzner [2022]), while crystallization refers to rich triangulation: the metaphor is of having so many facets of the data that it becomes like a crystal (Tobin and Begley [2004]). So, crystallizing a schema is the process of going from a vague idea about a dataset in your mind, to having concrete external representations of the schema that you can reflect on, share, and iterate, which then sharpen the internal schema.

To support the above, we present our own process of going from a vague research idea, expressing it within Teleoscope, and eventually crystallizing our schemas into visual artifacts that can be shared. We present methodological insights for designers who are developing ML-supported qualitative research tools using LLMs, as well as qualitative researchers who are currently developing meaningful definitions of rigour within this new field.

6.2.1 Thematic exploration with Teleoscope

Teleoscope is a previously-developed web interface for thematic exploration of large corpora in the range of thousands to millions of documents (see Figure 6.2) (Bucci et al. [2024], Lissandrini et al. [2019], Teleoscope.ca [2024]). Teleoscope allows researchers to create visual workflows to process and customize document similarity metrics using semi-supervised topic modelling techniques (El-Assady et al. [2019], Fariha and Meliou [2019], Nikolenko et al. [2017]). Thematic exploration is a term that is meant to reflect applying the values and rigour of thematic analysis to document exploration in large corpora (Nowell et al. [2017]). Teleoscope's workflows are meant to be externalizations of a researcher's schema so that they can check whether the machine model of the schema reflects their internal schema (Li et al. [2023]). Further, since Teleoscope is for collaborative workflows, the externalized workflows allows multiple team members to come to an agreement

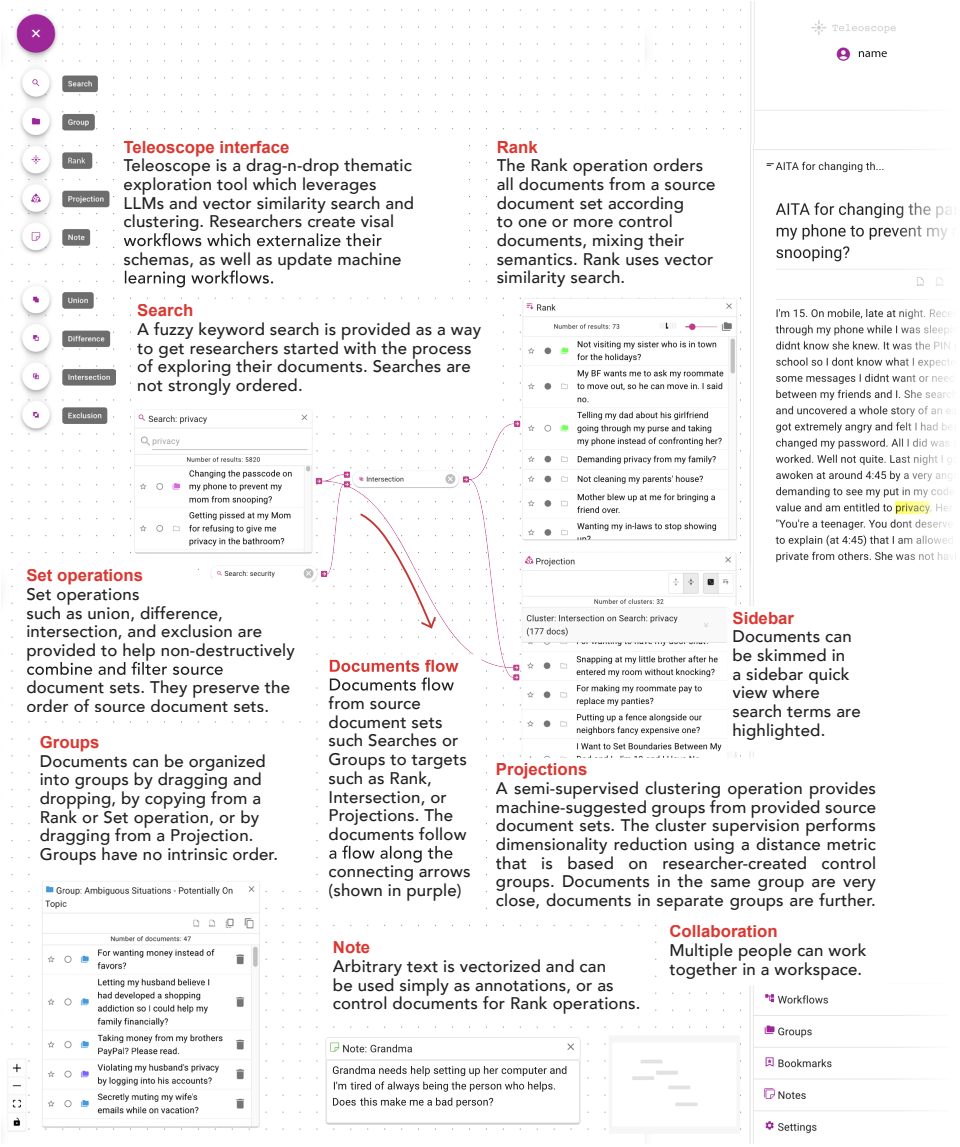


Figure 6.2: The Teleoscope interface.

about the theme as well as the process of how they came to the conclusion by tracing the workflows.

A simple way to think of the difference between a thematic exploration vs. analysis is the difference in inputs and outputs. A thematic exploration has a large dataset as an input, and the output goal is a set of grouped documents, associated visual traces that explain the provenance of the document groups, and annotations that develop the thematic justifications for inclusion/exclusion of documents. The exploration process operates at the level of document sets. However, a thematic analysis operates at the level of individual lines of text, typified by the hand coding process where documents are deeply read, interpreted, and related to each other. The output of a thematic analysis is a set of richly written themes supported by a system of coded and annotated lines of text within documents (Braun and Clarke [2022]).

As a theme forms conceptually, it is externalized on the Teleoscope interface as a group of documents where each document has been carefully considered. Then, a Rank operation captures the set of documents that we cannot possibly read, but have built up a complex enough network of interface operations and annotations to have a meaningful thematic explanation for the inclusion/exclusion criteria.

Our goal with this paper is to explore the process of schema externalization and sharing, and formalize the methodological approaches to support this new area of LLM exploration of large corpora.

6.2.2 Privacy/security in Reddit's AITA

We were interested in exploring privacy/security concerns within Reddit's Am I the Asshole (AITA) advice forum (Reddit.com [2024]). In the AITA format an original poster (OP) will ask a question about a situation in which they are unsure whether they acted like an "asshole," describe the situation, then ask the AITA community to vote on whether or not they were "assholes" in that situation. The AITA community will discuss in large comment threads while pronouncing concretely either "You're the Asshole" (YTA), "Not the Asshole" NTA, or "Everyone Sucks Here" (ESH). These votes are automatically tallied and included as tagged flair on the original post. Occasionally, the OP will provide further information, comment, apologize, or argue with the AITA community.

We chose this dataset because it provides an interesting explicitly normative account of relationship conflicts. The corpus is large, with nearly 1 million documents, which we filtered down to 300K usable documents by

removing posts that were removed by moderators.

Our original exploration topic was intentionally broad (Chen et al. [2018b]), looking simply for general privacy/security concerns that involved technology. Eventually, we settled on "Nuanced challenges to our current privacy model" as the first large-scale theme that we would explore. This became the nucleus of our schema crystallization: trying to decide what we actually meant by that.

At the moment, Teleoscope does not support nested document display such as with Reddit's comment threads. Since our thematic exploration is not a thematic analysis, our dataset can be constrained to the original posts.

6.3 Schema nucleation

Schemas are a concept in psychology that refers to a set of beliefs, attitudes, and emotional responses related to an archetypal situation, e.g., "driving." An unstated belief might be that "all drivers are not paying attention" which is related to a dismissive attitude towards drivers, and an emotion of frustration. If a driver in front were to take a few seconds too long before moving at a green light, the holder of the schema might externalize their schema with an explanatory statement by saying out loud that the driver "must be on their phone" whether or not that is true. After saying it out loud, a passenger may reflect and amend the externalized schema by saying "maybe they were just waiting for a car to finish turning."

Berret and Munzner introduce the concept of schemas as part of the sense-making process with data visualization (Berret and Munzner [2022]). Similar to the psychological concept of a schema, data visualization provides a method by which a researcher can externalize their schemas, allowing them to inspect, amend, and develop their internal schemas along with the externalized artifacts of the schema.

We define schema nucleation to reinforce the metaphor of a growing crystal for schema crystallization. A large crystal grows from a small nucleus. Similarly, expressing and developing a schema must start from a small entrance into the dataset. In our case, this was a standard fuzzy keyword search (see figures 6.4 and 6.3).

6.3.1 Schema Nucleation: Keyword Searches

Many qualitative researchers are formally trained in methodologies that manage keyword searches, including boolean search. A fuzzy keyword search produces a definite set of documents (see Figure 6.3). However, with a large

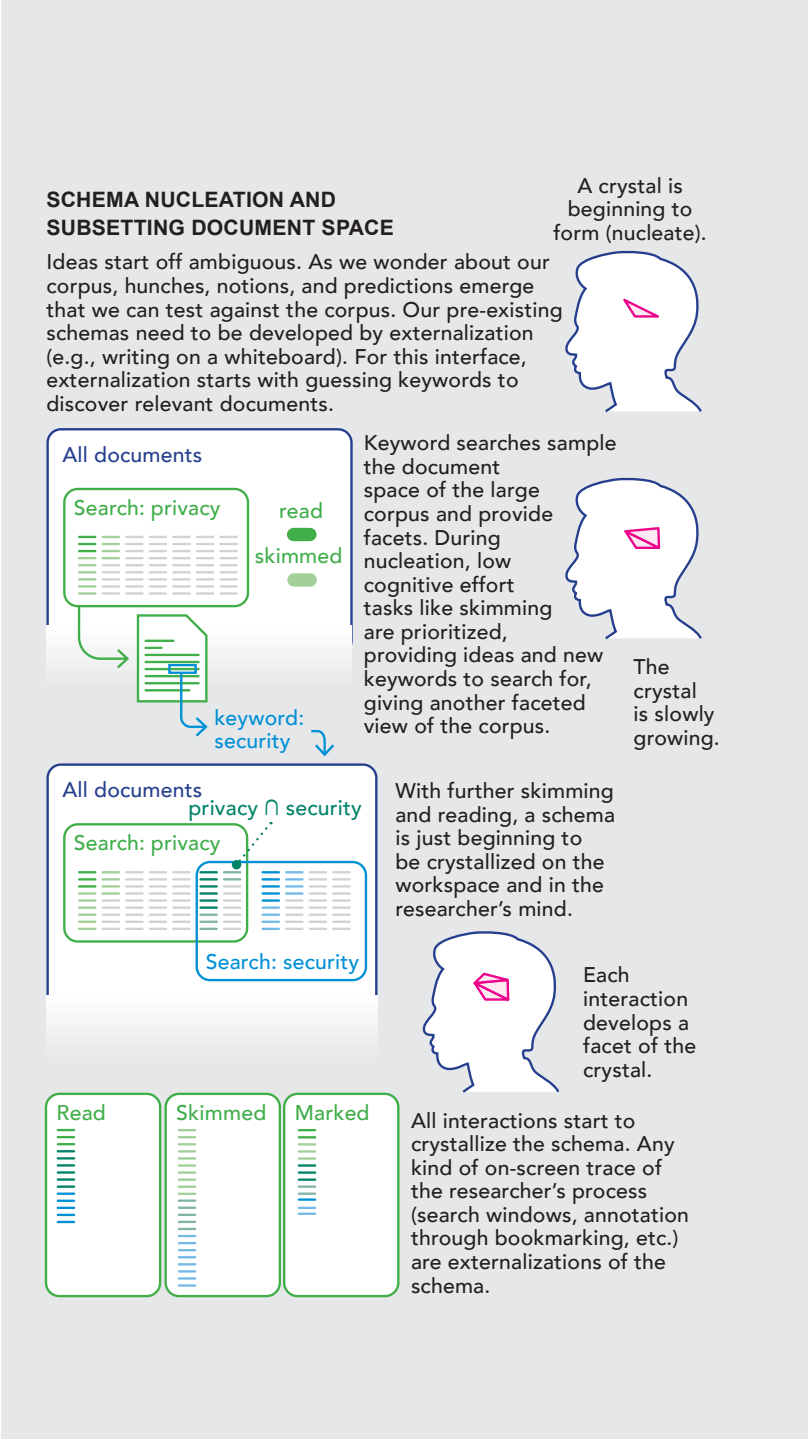


Figure 6.3: Schema nucleation.

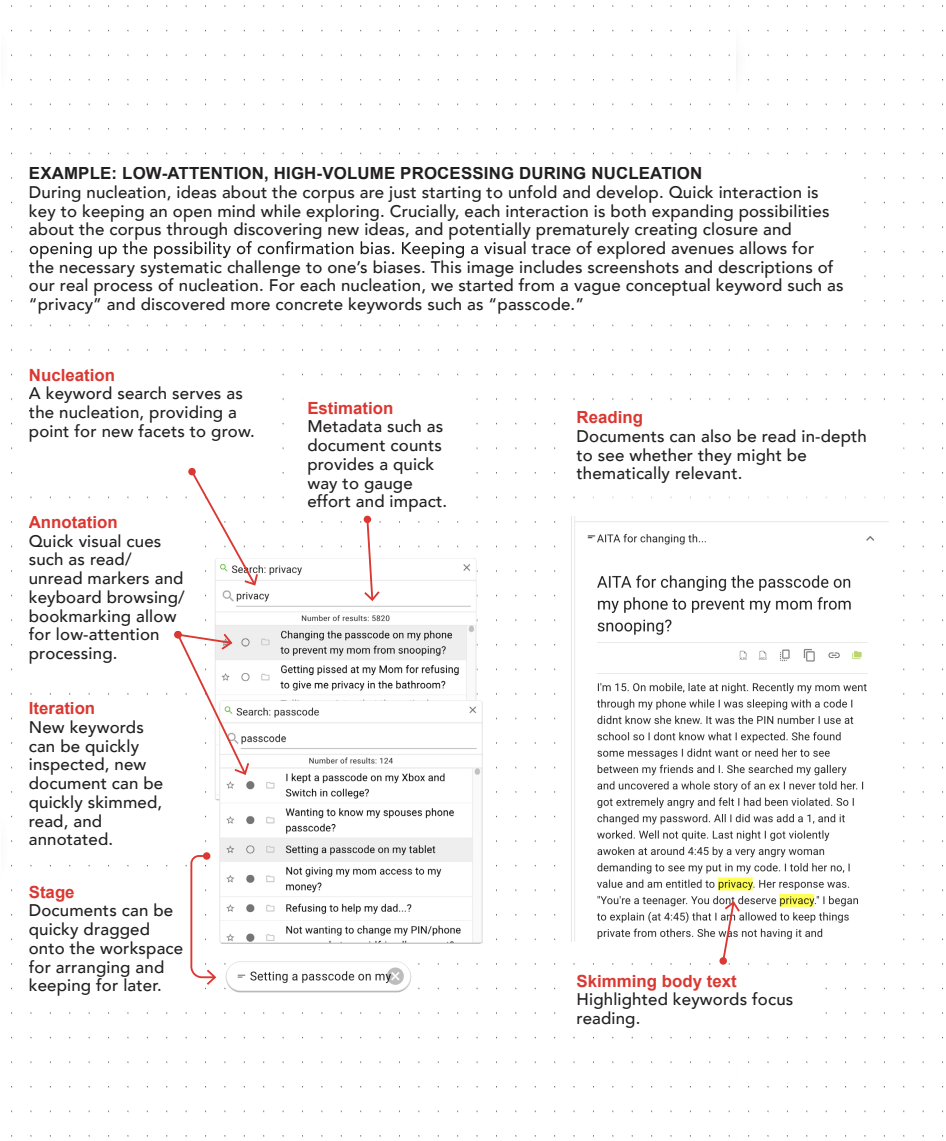


Figure 6.4: Schema nucleation example.

corpus, it is not possible for a person to read all the returned documents. Perhaps, one can skim over many titles, read a few documents in depth, browse through a few more to generate more keywords (see Figure 6.4). These first impressions comprise the schema nucleation phase: the researcher is just beginning to develop a point of view, but it is yet unformed. Every keyword search is a strict subset of the total document space, but it can also be seen as a facet, or view of the data. Every skimmed document is starting to shape the researcher's early theories of what the dataset does and does not contain.

Our Keyword Schema Nucleation Process

We began our exploration of the dataset with keyword searches for common privacy/security and technology related concepts, such as privacy, security, password, account. Immediately, questions of biasing our results through the choice of keywords arose. How did we know that we had the right keywords? How do we determine whether the posts captured by those keywords are appropriate? By skimming titles and reading through a few documents, we quickly discovered more concrete keywords such as "passcode" and "camera."

6.4 From keyword search to groups

After skimming titles and reading through many documents, we started to get a very broad and general sense of themes that were within the documents we had seen so far. For example, privacy brings up documents that were potentially relevant to our search, such as those about changing passwords, or reading through a partner's messages. It also brings up documents that are less likely to be relevant, such as using the toilet in private. At this early stage, we were not yet sure that any documents might be relevant or irrelevant: perhaps non-technology issues such as toilet privacy could shed light on technology issues. Documents that are interesting can be annotated, dragged onto the interface for later. We started to create groups of documents that captured vague ideas about undetermined themes. Importantly, although every interface action helped to develop our internal schemas, there was not yet any strong externalized commitment to the schema. This was a generative, playful, exploratory stage, rather than a stage with highly directed action.

With enough of this initial exploration, the interface's workspace becomes a visual artifact that serves as an external representation of the am-

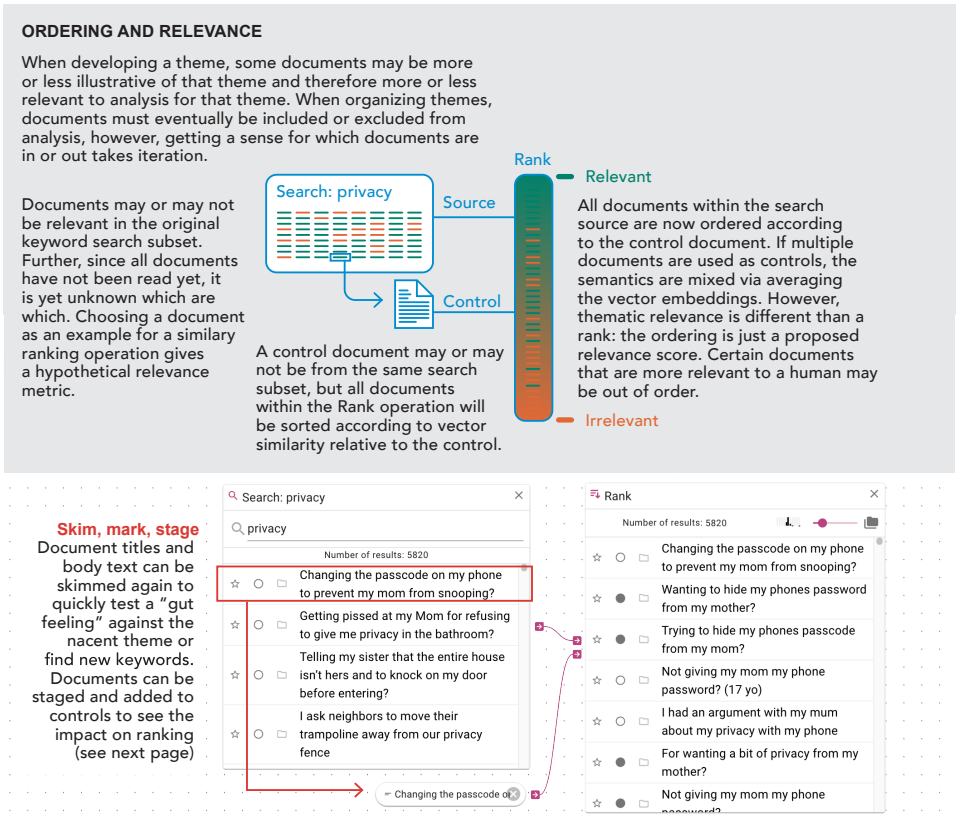


Figure 6.5: Ordering and relevance.

biguous internal schema. The workspace also can become messy, with staged and annotated items in no clearly meaningful arrangement (see Figure 6.5). To continue to develop the schema, a perspective needs to be imposed on the dataset, which grows a facet of the crystal.

To start to make sense of our keyword searches, we needed put them into some kind of order. With LLM embeddings, vector similarity search is an approximation of semantic similarity. However, there is a difference between one's schematic sense of thematic similarity and the sense of similarity in a vector similarity search.

Machine similarity is a statistical similarity, and there is no inherent discrimination between underlying word senses, interpretations, theoretical perspectives, or meaningful poetic connections. For example, in the figure on the previous page, the Rank operation is controlled by a document that is complaining about a mom "snooping." There is a semantic similarity that is captured with the subsequent documents referencing "hiding" but it does not capture the underlying sense of motive or discomfort that is implied in the document—these are our interpretations which are not necessarily present in the text, but instead in our own schemas. Imposing an order based on similarity ranking is a helpful first step, but we could not assume that the machine's proposed order would be the same as a human's thematic grouping. As we skimmed documents, we added new controls to better approximate the yet-undetermined themes. Rankings were updated to reflect the controls. We tested our "gut feelings" as we externalized by arranging and grouping documents before themes could be formalized in a concrete way. Eventually, we wanted to create a group of documents that were confirmed to be within a particular theme.

6.5 Determining saturation

Richness and saturation are two values that are used to talk about rigour during theme development (Braun and Clarke [2022], LaDonna et al. [2021]). They are defendable through rhetoric, but not measurable, which makes them difficult to mechanize within a computational system. Both terms refer to having dense meaning, richness through making connections to theory, between data points, and drawing in other sources; saturation through each new data point not providing a new perspective on the theme (see Figure 6.6 concrete example).

While thematically exploring large corpora, we would say that it is important to determine saturation because we want our themes to be rich.

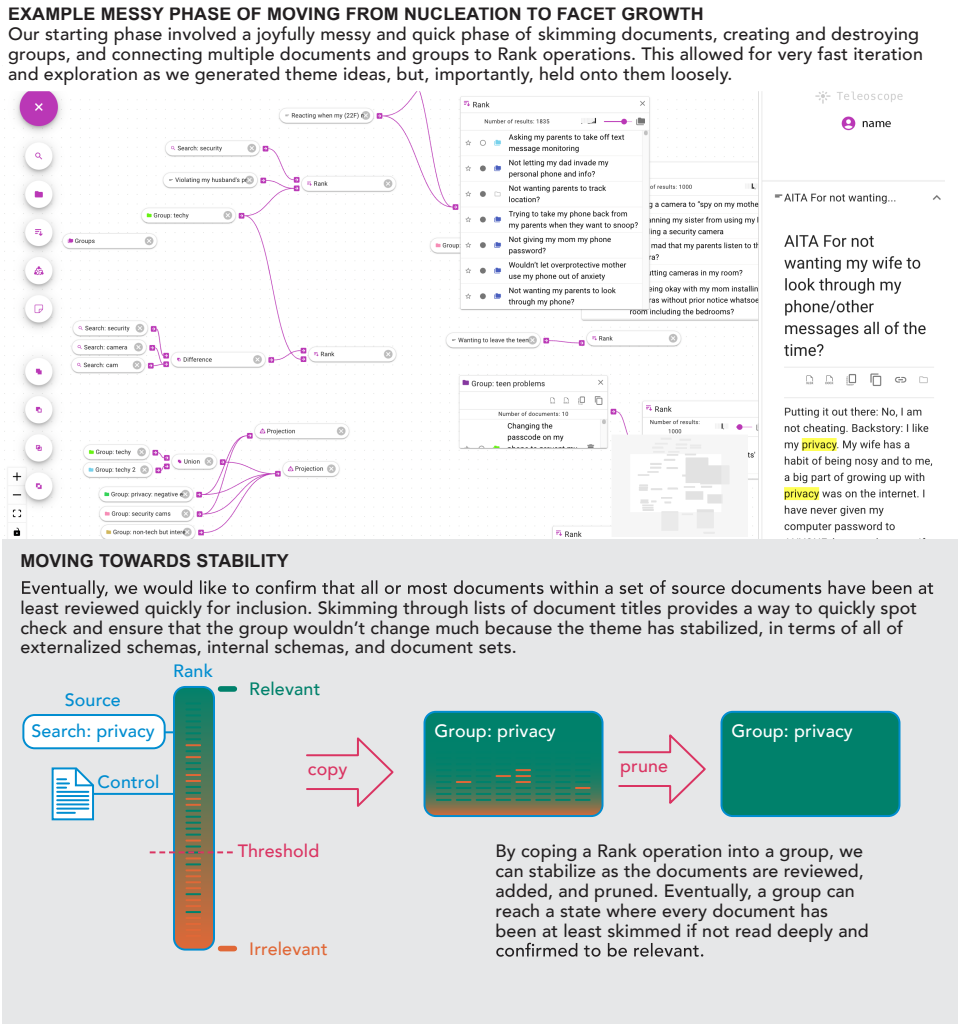


Figure 6.6: From messy to stable.

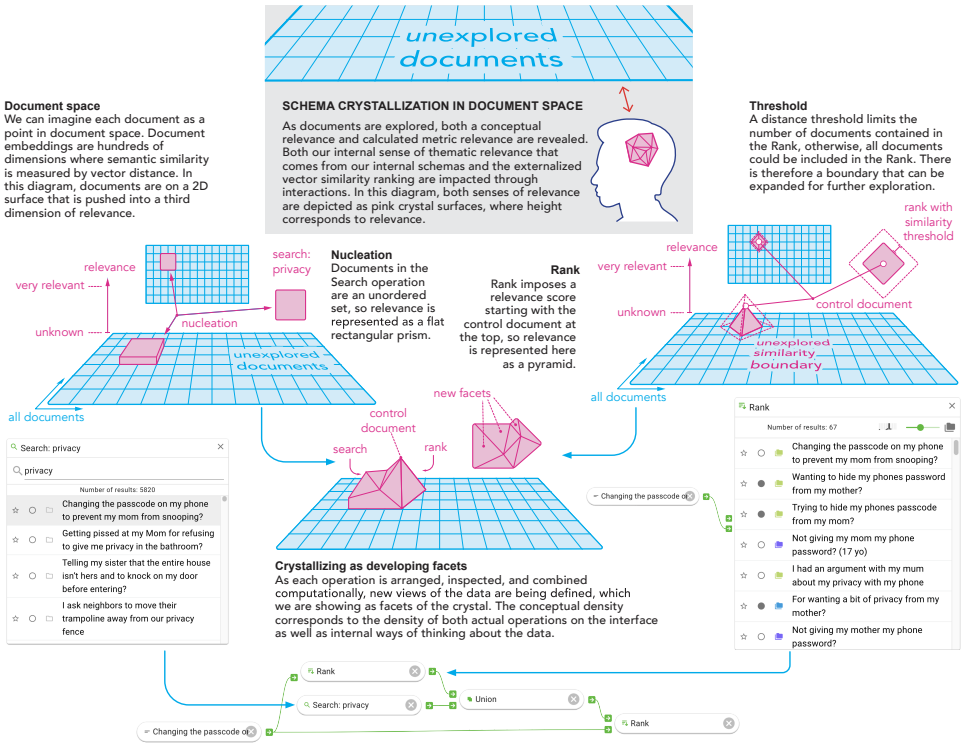


Figure 6.7: Schema crystallization in document space.

Remember, we are not doing a final thematic analysis with this method, rather we are trying to find the right dataset that will be able to draw out certain themes of interest. Therefore, we needed to find a way to define saturation in terms of using a system like Teleoscope.

Our approach is to liken saturation to a high level of crystallization (Tobin and Begley [2004]). Many facets are available to the researcher: Search and Rank operations have been iterated on many times, Groups have been annotated, broken into different thematic subsets, and new unskimmed or read documents are not often appearing. In this section we will explore what it means to determine saturation under the conditions of uncertainty that are intrinsic to large corpora.

6.5.1 Crystal facet density is equivalent to Saturation

With large corpora, it is impossible to actually read every relevant document. Some qualitative researchers will hire large teams of coders to achieve results

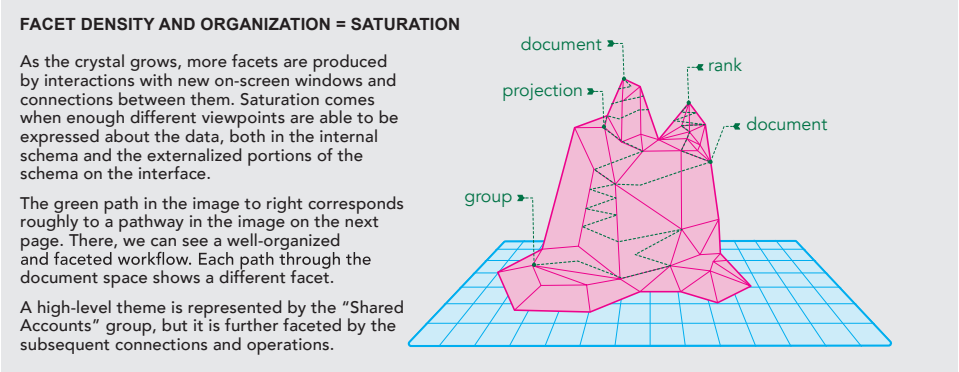


Figure 6.8: Facet density and organization.

that are not dissimilar to a mechanized result, because the enormity of the operation requires a systematic approach. Coders are given detailed instruction manuals and have team meetings to ensure interrater reliability. However, it is more in line with the values of thematic analysis to have individual interpretations that are shared and develop through collaborative discussion.

In our process of exploration, we found interesting example documents that became our nucleation points. Often, they were statistically similar to other documents that we found much less interesting, but not meaningfully similar. This was a surprising distinction.

Through our initial exploration, we found large groups of documents that we found to be too similar to analyze individually. For example, searching for wifi as a keyword produces many documents that are about fairness in paying for wifi, and the social difficulties of determining who has access to wifi. After exhaustively looking through every wifi document, it was clear that our dataset did not have any documents that included the wifi keyword and were about a radically different situation. Rich themes could be found within a thematic analysis, but further faceting within the wifi keyword search was not needed.

However, the thematic exploration was not yet done. Wifi is only one keyword. Through a Rank operation, we were able to find similar themes of family sharing of accounts such as Netflix, Spotify, and other streaming services. We made a list of keywords of unique streaming services with similar problems, similar to the qualitative researchers in the original Teleoscope paper (Bucci et al. [2024]). Like them, we reached a point of keyword saturation. Assembling all keywords gave a list of 13717 documents, which

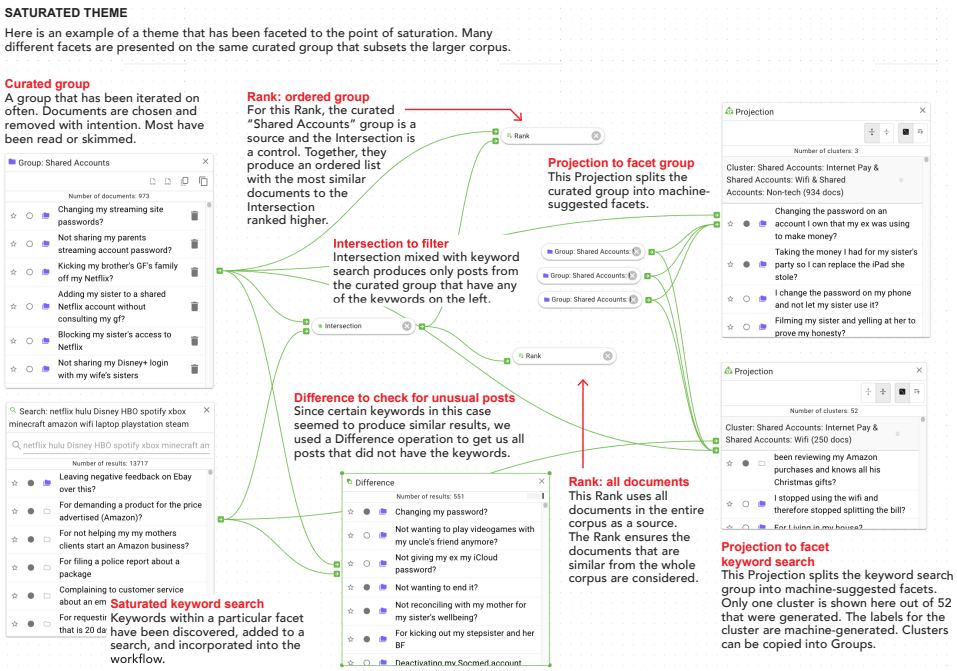


Figure 6.9: Saturated theme.

is too many to read. Further, even if we had captured a reasonable set of documents for this one theme, it had not yet crystallized, since there were not yet facets created.

Human-created groups indicate that the person feels that certain documents potentially belong together in a theme. When a few groups are created, a Projection operation can be effectively used to have machine-proposed groups (called "clusters") based on the human-made groups. This provides more facets through collaborating with the machine. Clusters can be dragged onto the workspace just like groups, where they can be refined, added to, and used for other operations such as Rank to find similar documents that were not in the original Projection source.

6.5.2 Incorporating Annotations and Arrangements

Annotation is an important part of externalizing schemas and developing themes. Arranging documents on an interface develops an external thematic understanding. Just like with affinity diagrams, arranging and re-arranging allows for quick, iterative mini-themes to emerge and disappear. Eventually,

themes congeal and need to be annotated to be properly understood.

Since Teleoscope allows annotations to be vectorized, this affords an interesting new mechanism of searching, which we will call searching by archetype (see Figure 6.9). As themes are annotated, common narratives emerge archetypally representing a theme. For example, "mom always snoops through my messages" might be a phrase that we would assign to a persona developed from posts about password sharing on mobile devices. Since they are also vectorized by the Note operation, they can be used as part of similarity searches. This is a new and interesting way of curating results. If a researcher can accurately take on the voice of an archetypal poster, they can find results that are thematically similar. Guessing and testing can help to refine the archetypal search results. Multiple archetypal phrases can be used. This is a novel search mechanism that is based in a creative arts practice that has essentially not be enabled with prior technology. Annotations can also be quotations. As example posts start to emerge, refining the vector similarity search using quotations allows for more on-theme results. This can be creatively combined with archetypal searches as well as document similarity search for further semantic mixing.

6.5.3 Signposts are differentiating examples

The more we worked with our documents, the more we found that we had (1) documents sets to analyze in large batches; (2) documents sets to confidently ignore; and (3) special example posts that were "exceptions that proved the rule," warranting special focus and attention—signposts. Signposts were closest to our internal schemas, but difficult to express as external workflows. They became key documents to determine saturation and major targets for search. By finding these hidden gems, we anchored our crystallization efforts.

An example is "AITA for secretly muting my wife's emails while on vacation?" This post (see Figure 6.11), along with other signposts, helped us to solidify our thematic exploration around "ambiguous privacy/security violations that challenge the accepted value systems of current privacy models." However, this was an extremely difficult theme to look for, only discoverable through carefully reading most of a Rank operation. That is because the content of the post is a nuanced problem about searching through a partner's email, which rarely surfaces posts about nuanced situations. There is no nuance detector built into vector search. Despite searching using keywords, Rank operations, and archetypal searches, there was no automatic way to bring up posts that were similar in underlying nature. This illus-

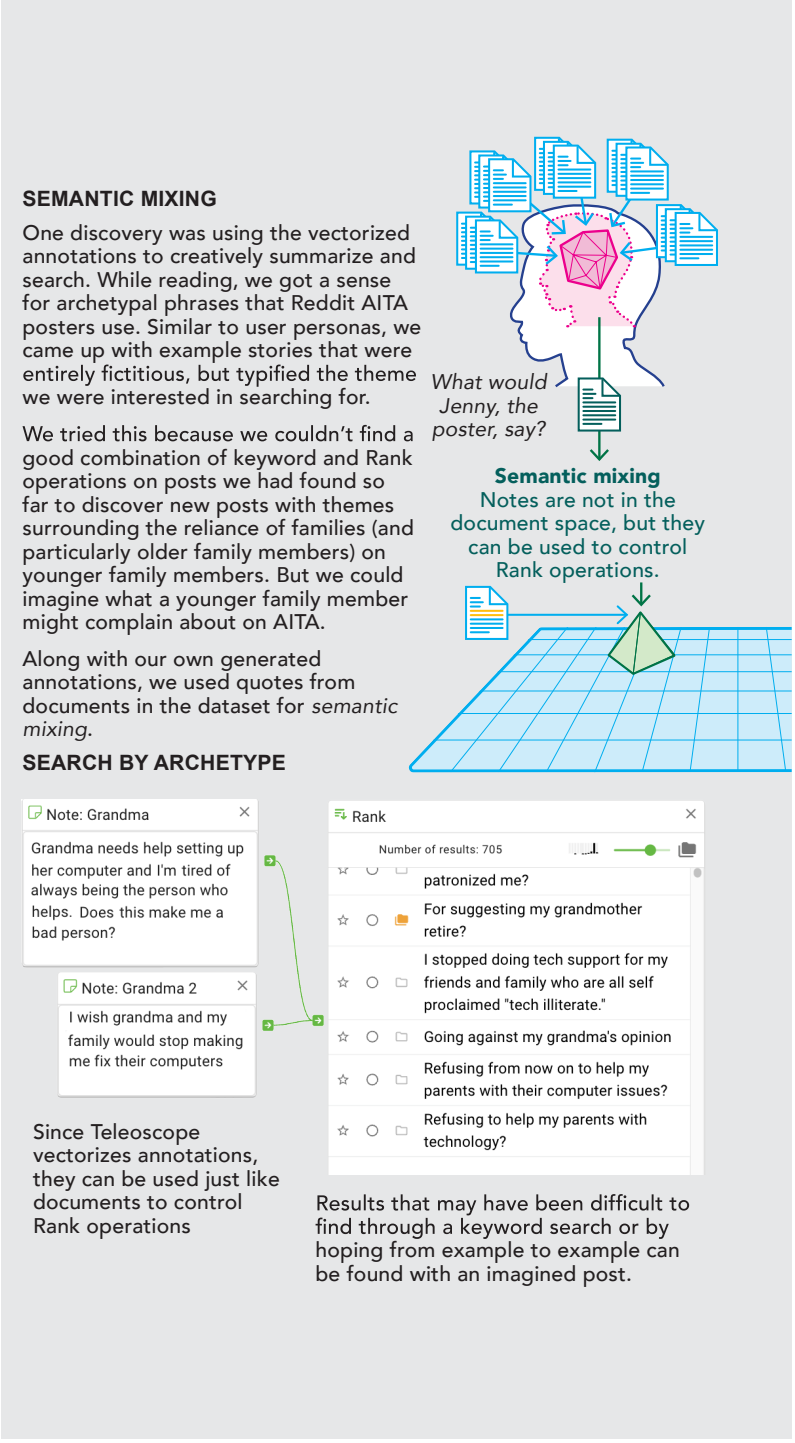


Figure 6.10: Semantic mixing.

trates the fundamental difference between what a machine can easily do with a vector search vs. the value of a human interpreter.

Our interpretation of the post is that there are conflicting value systems at play. On the one hand, the husband is clearly violating a privacy norm and controlling his wife's actions. On the other hand, he believes that the is doing her a favour, and may very well be improving both of their vacation experiences. This level of nuance is essentially not possible to discover with a vector similarity search.

Our saturation goal therefore became to find the most different examples of nuanced situations such as these. Semantic similarities gave us a recursive process, even within categories of interest breaking them down into large positive examples, negative examples, and special examples. For us, saturation came when further faceting could not meaningfully differentiate groups of documents; as a corollary, we could not find any further examples that could signpost differentiating factors of our themes.

6.6 Results of exploration

Results of our exploration are document groups supported by workflows and annotations. A thematic analysis would be needed to report on themes. However, it would feel unsatisfying to leave our results entirely unreported for this pictorial, thus we have summarized our findings here. Further, the Teleoscope platform allows people to create and share workflows.

Nuanced Account Sharing vs. Snooping, Spoofing and Spite

Our common privacy models do not consider the complex interplay of relational values evidenced by the marked difference between the way account privacy systems are designed vs. how the accounts are actually used. In this thematic exploration, we have surfaced exemplar documents from a large corpus of advice forum data that establishes themes within account sharing where our privacy models fail to meet clear relational needs.

Explicit account sharing models have been established for streaming services such as Netflix or Spotify. Through our exploration, we found that any shared resource such as wifi or even physical devices such as laptops had similar themes in terms of relational conflict. We have organized these document sets in terms of the nature of conflict and identified archetypal documents which represent the larger sets. Cross-cutting themes appear to be in terms of the method and justification for removing someone from an account, including intimacy, socio-economic status, and relational rupture.

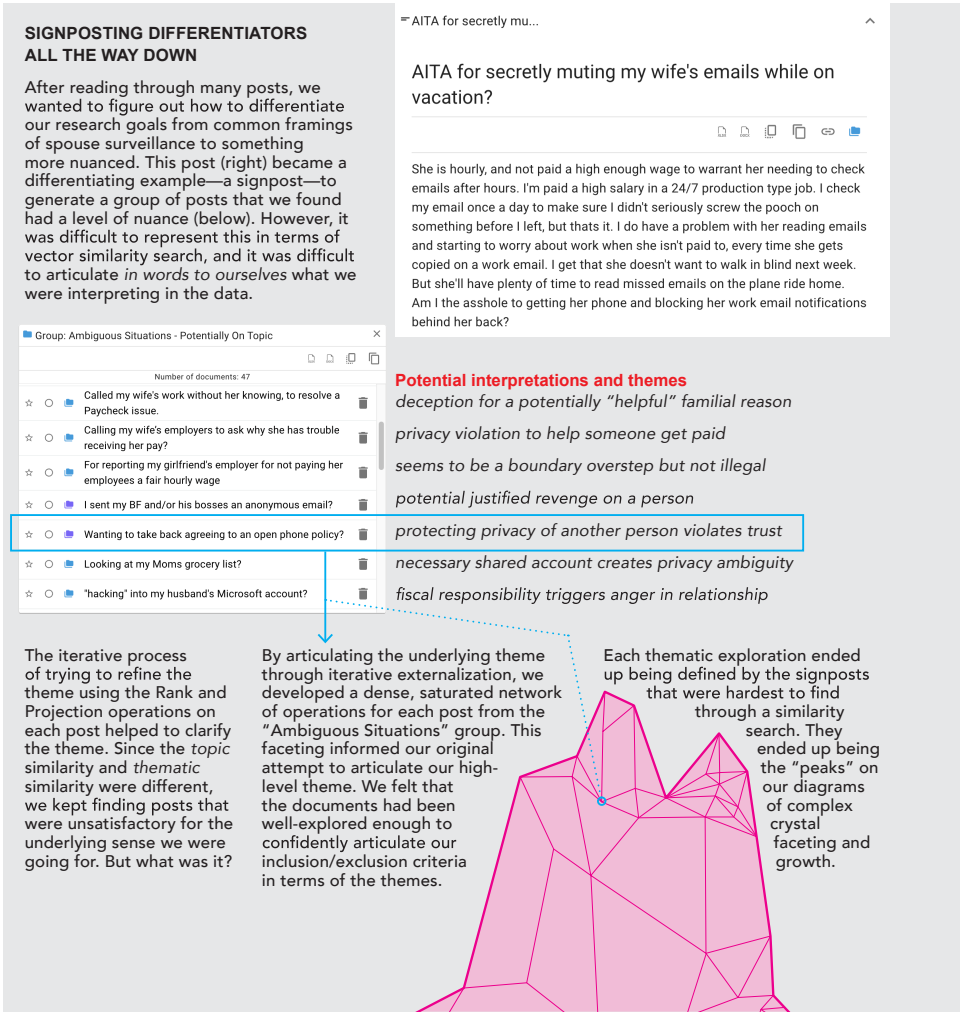


Figure 6.11: Signposting.

Other explicit sharing models (e.g., social media sharing, Dropbox or iCloud) enable sharing with permissions, but since the nature of the data is user-generated, it can produce conflicts over social expectations of privacy and trust. Themes included power imbalances such as adult children managing parental data, peer conflicts with unexpected transparency or lack of privacy such as with adult siblings seeing each other's assumed private data via a parent's device, or expectations around the privacy of social media communities and data permeability.

However, the bulk of our analysis focused on nuanced conflicts within intimate relationships regarding privacy models that assume a single-user-single-account, but in reality, have multiple stakeholders due to intimate interdependence. We found email, bank, and social media accounts as well as user enviroments such as browsers potentially were assumed within intimate relations (e.g., partners, parent-child) to be shared as part of establishing trust (e.g., through sharing passwords). Paradoxically, actually acting on an agreement such as an "open phone policy" was often seen as a violation of trust. Security violations such as spoofing were considered to be justified when other greater interdependence values were violated (e.g., non-payment of bills, interventions into work disputes). Relatedly, access to other people's private information via chat messages was not necessarily considered to be a violation unless enacted in a way deemed unfair or indicating inappropriate power over somebody.

Excluded from the main analysis, but considered for evidential counterpoints were camera surviellance in shared spaces, trust violations with regards to pornography, cheating, location tracking, and privacy violations that did not include computer technology.

6.7 Reflections

In this pictorial, we have presented a central metaphor of schema crystalization that articulates the process of making sense of large corpora using a tool that leverages LLM embeddings and vector similarity search. We have theorized about a methodology called thematic exploration that justifies document inclusion/exclusion into document groups that are accompanied by retraceable workflows and accompanying annotations. We reported on our own experience with this methodology, developing analogies to thematic analysis within this new area of LLM-enabled qualitative methods (see Figure 6.12). We hope these can articulate ways of dealing with the immensity of qualitative analysis within large corpora and operating within the uncer-

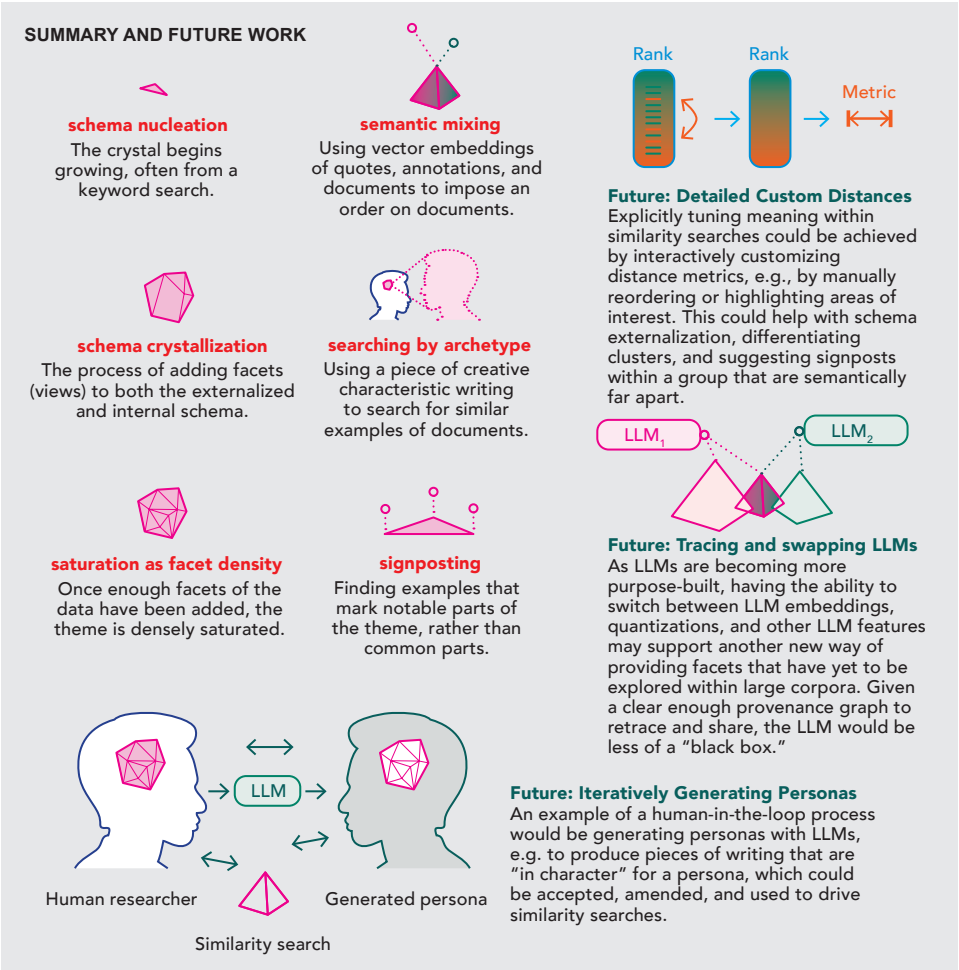


Figure 6.12: Reflections.

tainty of black-box approaches to LLMs.

Through our experience, we believe that future work can incorporate LLMs more deeply into qualitative methodologies provided they are not superficially used as black boxes that magically produce summaries, themes, or other writing (see Figure 6.12). Our own process involved finding limitations of LLMs in terms of producing results that rely on tacit and emotional information, interpretation, and implication We speculate that these subtextual elements would be difficult for LLMs to encode directly for, however, may be revealed through human-in-the-loop interaction as interfaces such as Teleoscope are developed.

For example, the process of semantic mixing and creating archetypes would not have worked "out of the box" with an LLM that simply summarized or tried to prioritize relevant documents without human input. That required a creative iterative process to draw out the "hidden gems" from the dataset. Even though we had externalized portions of our schemas through the crystallization process, the LLM does not have direct access to our internal schemas and is unchanged by the process—whereas we are. By externalizing our schemas, our unstated beliefs are now available for us to reflect on and share with others. As a result, we might update our beliefs about our corpora and the themes therewithin. Once we have externalized and inspected our schemas, we have also changed our perspectives on the world.

Chapter 7

Conclusion

In this dissertation, I have presented an account for making meaning with computational devices, both from an *embodied* and an *extended* perspective. The embodied perspective focused on theoretical approaches to sensing and displaying emotions to both support and critique work that I had done with others to make sensing systems for human emotions, and robots for emotional interaction. The extended perspective focused on the design of *Teleoscope*, a system for making meaning from large datasets by interacting with machine learning workflows. With it, we translated values and rigor from thematic analysis into machine learning-supported data curation which we call thematic exploration. Summarizing our experience using Teleoscope for analysis, we presented the process of schema crystallization to describe how the meaning-making process can look with computational support.

To frame it another way, this dissertation explored the subjective experience of meaning in a computational context from the ground up: from sensation to cognition, from affect to schema, and from individual to collective. I have found that by trying my honest best to represent difficult concepts using machines with the precision of computation but the limitations of practical engineering has been a surprisingly poetic journey. In this chapter, I will briefly summarize the major concepts and themes that underlie the dissertation and reflect on the insights that the work presented here can give to those concepts.

7.1 Mixing methods: Quantitative vs. Qualitative

Due to the computational focus of this dissertation, all of the work presented here is based on logical and mathematical processes at some level. A fundamental assumption is that subjective experiences are physical processes, and as physical processes, they can be detected and represented in a computer. There is a kind of computational realism hidden in the dark corners of this work, that is, something exists only insofar as we can compute it. Almost

as if only we could make better computational representations, we would better approximate reality.

However, my experience attempting to study subjective phenomena has been that the more measurement I attempt, the less the phenomena is like the experience I am interested in representing. In trying to measure emotions very rigorously, I found that it didn't really make sense to "measure" emotions. In trying to represent meaning, I found that meaning would be fundamentally changed by the process of representation.

I remember my qualitative research professor was profoundly confused when I proposed to use a "mixed methods" approach to make emotion scales better represent subjective experience. I can only hope that I have gained a sliver of the wisdom necessary to be confused by the idea of making the qualitative and quantitative meet.

Another short story on that same problem was when I presented at the ACII 2019 conference for my work in Chapter 3. A conference attendee exclaimed that my work was nonsensical because "emotions are social signals." How strange! From an engineering perspective, all signals should be processable, right?

My reflections on this are answered by practical limits to computation as bounded by complexity. Even if I would like to believe that, at some theoretical level, if only we knew all of the positions and velocities of all particles involved in a system, we could predict the future, my practical experience with robotics quickly taught me that simulations and models very quickly degrade with the number of variables and iterations over time. And some of the most unexpected realities are true, i.e., we can use GPS to find someone on the face of the planet within a few meters, but precise robot positioning within a room is still a very difficult problem. Scale differences actually matter. My imagination of the theoretical feasibility of the problem has been very different than the practical reality of implementing a solution. And "better technology" isn't always the solution: one of the most effective positioning systems for highly precise robotics continues to be a simple push-switch. The idealism of Silicon Valley "we can hack anything" is exciting, but ultimately a marketing facade.

The gap between quantitative and qualitative is so large that it is practically unbridgeable (although of course it is possible to triangulate). Experimental claims from a particular experimental setup are often different than the ways in which the study implications are talked about. Depending on your research questions, you *can* use quantitative methodologies to talk about subjective phenomena, but you will miss important parts of the phenomena if you do so.

In Chapter 3, I eventually realized that the critique from the ACII'19 conference participant was that social signals involve a large set of hidden variables that would be practically impossible to account for. The phenomenon of a "social signal" happens in such a complex system that it becomes impossible to compute reliably, particularly because the data collection methods collapse a large dimensional space into a small enough dimensional space that humans can reason about it. For example, in studies where we use word labels to describe emotion states, if indeed our statistics are sound, we are only capable in good faith of making probable claims about word labels, and not very much about emotion states, because of that dimensional collapse.

For example, I would not use the PANAS anymore. The PANAS is a circular measure: it is a measurement tool that measures a participant's understanding of a word relative to a common cultural understanding. The PANAS is a likely set of words that people from a particular culture may use when asked to rate their experience using an affect grid. If we were to analyze the probability that any particular word from the PANAS was being accurately used, it would be staggeringly low. We would have to calculate the probability that the PANAS study participants had similar understandings of our words to our own participants, never mind the precision of the words and their ratings and the alignment to purported inner states. In some way, an experiment designed to use the PANAS is relying on my current participants being very statistically similar to the original PANAS participants. Which is a dubious assumption, especially since many of my participants are from different cultural groups, and we might have to teach them what the words "mean."

It isn't that the words from the PANAS don't represent emotional states—they must to some degree—it's that the process of studying emotion states using word lists requires these summarized, externalized symbols (the words) to be concretized (the measurement), when human experience is an ever-unfolding process (the phenomena). It's not that there aren't ways to quantify qualities, it's that the static symbolic representation captured in a word can't capture the fullness of the category of emotional experience that we call "qualitative." So, the interpretive methodologies used in qualitative methods don't typically use quantitative statistical methods because of this flattening effect. When we use quantitative methods, we are necessarily throwing away the richness of individual experience. Perhaps for good reason—some studies don't require richness to have an effective research output—but perhaps not.

My conclusion is that it is certainly possible to use mixed methods to

research a single phenomenon, but that the triangulated results may be difficult to combine in a statistically consistent manner. That is, mixing methods doesn't bother me, but the "mixing" needs to happen on the rhetorical side, i.e., during interpretation rather than measurement. This is why Telescope ends up being process-oriented rather than output-oriented. The meaning system needs to include the live, on-going interpretation of the human being to be valid; the artifact is just a trace of what the human was willing to concretize along the way.

7.2 Machine Learning vs. Cognition

During my time studying natural language processing (which has been at least a decade, maybe longer), I have seen the field change from a tentative interest in neural networks to being totally dominated by them. When I started studying them, the question was "will larger language models outperform old-fashioned theoretical linguistic models?" and the joke was that Google's prediction models got better every time they fired a linguist. My own understanding of machine learning has grown as I have moved from largely using classifiers to largely using LLMs.

Due to the name "neural network," there is a common question as to how effective our machines are representations of the human cognition system. With the incredible success of LLMs, it seems like we have cracked the nearly 70-year old problem of the Turing Test wide open. However, I am less optimistic about having found a general AI than the typical marketing of LLMs seems to like to prefer.

First, the same phenomenon of drift is happening to LLMs just like with physical sensor models, sometimes called "hallucinations," sometimes called "data drift," sometimes called "prompt drift." Just like with sensors, I think there will be some excellent engineering tricks to reduce drift, but ultimately some problems may come up against fundamental limitations. Again, we can track people's phones via GPS with incredible precision, but we can't actually get centimeter-level precision from autonomous robots in a room. The drift is too much and we need to use other techniques.

Teleoscope is a way to try to manage this problem from a practical HCI perspective. LLMs have limitations because they are not embued with the sum total of our experience, and could never be. The Teleoscope design insight is to provide a continuous, interactive way to modify human-understandable inputs to produce human-evaluatable outputs. It helps people decide what they want, with the understanding that that will

shift through interaction.

In Teleoscope, "thinking like you" means that the machine learning model will output document groups that are similar to unstated and subconscious thematic understandings that are felt, rather than understood. In some ways, it operates on subtext by using text encodings. The assumption is that the subtext is encoded to some degree in associations, and that the dimensional reduction will pare away the associations that are not relevant to the theme being explored.

As I have said in the dissertation, my working understanding of machine learning models is that they represent a lower bound in terms of the complex structure of cognitive processes. They provide insight into how our cognitive structures work in this sense, but it is a mistake to imagine that our biological neural networks are equivalent to machine learning neural networks. There is surprising and useful functional overlap that should be studied, but the limitations of the metaphor are very important to recognize.

For example, the mechanical metaphor was the most commonly used understanding of cognition up until the advent of artificial intelligence and search algorithms. If the brain is like a machine, we now know it is not like a machine with gears and cogs. But it is also not like a search algorithm. Chomsky grammars helped us analyze language using computational metaphors based in graph search, and there was great effectiveness in it for a time, but the metaphor is comparatively not very useful with the advent of machine learning. As just one example of a possible critique of the current metaphor, if our biological neural networks are "like" machine learning neural networks, then they must be much more temporally embedded than our machine learning neural networks currently are. My instinct is that something else will replace this metaphor soon enough.

The helpful takeaway is to use the complexity of machine learning neural networks as ways to complicate our current models and understandings of human cognition. We can use the power of interactive systems to model processes even if the models are almost entirely "wrong" in terms of describing the actual functioning of the human cognitive system, but, if my own experience attempting aspects of this is instructive, we will gain great insights from the attempt. This is more art than science; more poetry than engineering.

7.3 Saturation and Information Power

Teleoscope offers a way to visually represent saturation and information power by using multiple examples to build themes of inquiry. After reflecting on the design of Teleoscope, there are many ways that we could extend the work to better encompass the concept of information power.

With the *wifi* examples from Chapter 5, we found that many *wifi* problems felt identical from the perspective of account security. Similarly, they also felt identical to many other account sharing problems, such as with *Netflix* or *Spotify* accounts where friends and family members were fighting over access to the account and equal payment. In other words, each document had low information power, but high *statistical* power.

From a statistical power perspective, we could make claims about the prevalence of a phenomenon within the corpus, or estimate the prevalence in a population based on sample data. These are interesting quantitative measures, but not qualitatively useful.

From a qualitative perspective, it would be more useful to find both typifying and differentiating examples. This is what is captured by the concept of *information power*. For many *wifi* problems, the interesting qualitative thematic curation action is to find the *essential* or *typifying* document, but also the *essential outliers*. From an encoding perspective, the typifying document might be most similar to most other documents, but the *essential outliers* are not. They often have content similarity, but subtextual dissimilarity.

We referred to both essential outliers and typifying documents as nucleation points, since they often "grew" a schematic crystallization of a theme. The example from Chapter 6 was of a husband who secretly muted his wife's emails when they were on vacation. Despite extensive searching, it was difficult to find any other documents that "felt like" this document in terms of the moral ambiguity of the situation. We might call this an "essential outlier", since it was unusual but said something that felt *perfect* about the theme. In other words, this document had high information power, and low statistical power.

7.4 Positive Implications for HRI and HCI

A large part of this dissertation involves working through my personal difficulties with taking a cognitivist approach to emotion recognition, when it seems that the phenomenon is more complex than is typically construed in

HRI/HCI research. However, there are many positive takeways. Here, I outline a short vision for the future of affective computing in HCI and HRI.

From a design perspective, focusing on designing for meaning systems allows us to apply methodologies appropriate for meaning, rather than attempting to construe design research in terms of experiments or a facade of scientific generality. For example, I would personally find more value in a study about robot emotions where a researcher worked with a highly-trained puppeteer than attempted to construct an experiment that demonstrated the universality of the robot's emotion display. Deciding when a project is legitimately more speculative fiction or performance art than science would help save effort in trying to fit the work into the wrong box.

However, deeper than that, I believe that the theory of constructed emotion (TCE) provides a fascinating model of emotion that we can incorporate into our design work. In fact, Fridman, Barrett, et al. provide a fascinating example of applying TCE to a real-world problem of analyzing high-stress police activities and providing recommendations for training (Fridman et al. [2019]). They argue that understanding emotions as *constructed* implies that training procedures should focus on developing interoception, or a person's ability to focus on their awareness of the meaning of their felt bodily signals. This can help the person intervene on the construction of an emotion in real time.

The inventors of Biomusic (Cascio et al. [2020]) may agree. They have developed a system for translating body signals into music to support neurodivergent people in becoming more aware of their internal emotional states. Rather than the system labelling the signals with emotion words, with the help of a therapist, the participants use the signals to help narrativize their own experiences. In other words, the technology is supporting someone in improvisationally understanding themselves, rather than trying to make decisions for them, and maintaining the focus on human-human interaction.

My vision for HRI/HCI research that wishes to work with human emotions follows the same pattern. Recognizing the design implications of TCE would mean focusing therapeutic interventions on introspection or interoception (such as with Biomusic). My own personal experience of successful emotional intervention technologies in health care is that devices that flexibly support improvisational play are more successful than prescriptive systems.

Examples of prescriptive systems are typified of by a genre of HRI/HCI intervention: *stress detection in car systems* (Siam et al. [2023], Paredes et al. [2018], Zontone et al. [2019]). The system purports to have a full understanding of an affective state, and occasionally prescribes an intervention.

I am not able to evaluate whether such a design makes sense for a high-risk high-control environment like driving cars. But I was once asked by a research manager at a mining company whether I could build an EEG-based alertness detector into a hardhat for drivers of their large mining trucks.

My answer was, "yes, probably," but the request made me distinctly uncomfortable. What does an alertness detector for a driver do? What does a stress detector for a driver do? There's an implied intervention that to me seems to perpetuate whatever problem is trying to be solved by incomplete measures. If your drivers are falling asleep at the wheel, the ethical solution isn't to keep them awake through a technological intervention, the ethical solution is to create a system where they don't have to be worked to the point of exhaustion. Our technological solutions are not detached from their social contexts. At some point, the stress detector just becomes another monitoring system, adding to information and notification overload [9].

If we understand emotions as systemic, emergent homeostatic processes—as TCE does—the design implication is that no one intervention is going to create a shift within the emotion system. Instead, we should be focusing on designing technologies that do not prescribe as much as help a person describe and become aware of their own emotional status. My suggestion would be that making effective relational, improvisational and introspective technologies will get closer to producing *effective* affective interventions.

7.5 Extended and Embodied Cognition

The final reflection I have is on extended and embodied cognition. I have become a strong believer in embodied and extended cognition, and the aforementioned machine learning metaphors (coupled with a computational graph understanding) have helped me develop an intuition for why that might be.

Imagining the brain like series of neural network layers arranged in a very complex multi-layered weighted graph, it would make sense that certain areas of that graph would be only possible to activate if and only if other disparate parts of the graph are activated. Concrete and funny examples: why can't I remember a song lyric without singing the whole song sometimes? Or why can't I remember where my keys are unless I retrace my steps?

If we think of memory as entirely "in the brain," this question becomes

[9] A 2018 study showed that sometimes having more information can *impair* decision-making for nurses and baby care (Van Kollenburg et al. [2018])

"why can't I just run my search algorithm effectively?" which, I believe, is the wrong question. Rather, there must be a particular pattern of neuron activation that is necessary and the only way to go about activating those neurons very well may be by going through a process or with particular sensory input. We imagine the brain to have total access to itself, but, actually, the cognitive graph is very likely highly partitioned.

Hysteresis is actually a key working part of this system, exemplified by how touch sensation works (for certain signals). We eventually do not notice the pressure of our clothes on our bodies: why? Partially, there is a "software" reduction in attention, but also there is a "mechanical" reduction in the signals being sent to the brain due to hysteresis. There isn't enough neurotransmitter left due to the high frequency of touch signals from your cloths on your body, so signals are sent less frequently as before, therefore the perceptual signal is reduced.

The "embodied" part of extended and embodied cognition means that the cognitive system, typically though of as a brain or a computational system, exists in a body, and that the body is a part of the system, not external to it. A shocking thing that a biosimulation professor once said to me is that muscles are just neurons that move (they transmit action potentials and also contract). Further, he claimed that there are enough neurons in the optical nerve to see that it does some amount of "data preprocessing," that is, the "brain" does not do all of the computational work while the "body" does none. The closed feedback loops that exist in the muscles make it analytically possible to consider that the muscles are doing some amount of data processing as well (although it's often not thought of like that).

The "extended" part is to realize that the external world is actually also part of the cognitive system. This is non-trivially true, i.e., humans both require stimulation from the external world to perform cognitive processes as well as *identify with* parts of the external world. This is analytically counterintuitive but experientially normal: you probably say that someone hit "you" if they hit your car, and in a lot of ways, that is truly your experience of it.

The question that is often asked then is, ok, if all that is true, which part is the "me" part? Which part is the "experience" part? It really doesn't seem like my muscles are conscious, and people lose limbs all the time and still have subjectivities, and certainly my car is not really, truly part of me. In fact, the opposite direction of elimination is also true. You can lose huge parts of your brain—or even have it split in half—and still basically function and identify "normally." It seems that all the brain really needs to have an illusion of experience is to be able to rationalize causality.

I can't answer questions of consciousness, although they are fascinating. In the sense of having an illusion of experience, identifying with our external environments, and relying on the world to put our brains in a particular state of activation, I think there is a non-trivial and very real sense in which our devices and programs (such as Teleoscope) are indeed parts of our cognitive system, particularly while we are using them. I feel quite confident in claiming that they are part of our meaning systems (as written in the dissertation) but it also seems to be both subjectively and ontologically correct to some degree. So, whether the computer "thinks like us" is hard to determine, but we certainly have incorporated the computer into the "way we think." And to me, that's fascinating.

Bibliography

Ralph Adolphs, Daniel Tranel, S Hamann, Andrew W Young, Andrew J Calder, Elizabeth A Phelps, Al Anderson, Gregory P Lee, and Antonio R Damasio. Recognition of facial emotion in nine individuals with bilateral amygdala damage. *Neuropsychologia*, 37(10):1111–1117, 1999.

Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182, 2014. doi: 10.1109/VAST.2014.7042493.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020.

Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial Intelligence Review*, 56:1–81, 2023.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Fooling Neural Networks in the Physical World with 3D Adversarial Objects. `https://www.labsix.org/physical-objects-that-fool-neural-nets/`, 2017. Accessed: 2024-10-14.

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

Rajiv Badi, Soonil Bae, J Michael Moore, Konstantinos Meintanis, Anna Zacchi, Haowei Hsieh, Frank Shipman, and Catherine C Marshall. Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 218–225, 2006.

Kaveh Bakhtiyari and Hafizah 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.

Barnaby Barratt. *The emergence of somatic psychology and bodymind therapy*. Springer, 2010.

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

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

Lisa Feldman Barrett and James A Russell. *The psychological construction of emotion*. Guilford Publications, 2014a.

Lisa Feldman Barrett and James A Russell. *The psychological construction of emotion*. Guilford Publications, 2014b.

Elizabeth Bates. *The emergence of symbols: Cognition and communication in infancy*. Academic Press, 2014.

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

Katja Battarbee and Ilpo Koskinen. Co-experience: user experience as interaction. *CoDesign*, 1(1):5–18, 2005.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

Peter Berger and Thomas Luckmann. The social construction of reality. In *Social theory re-wired*, pages 110–122. Routledge, 2016.

Charles Berret and Tamara Munzner. Iceberg Sensemaking: A Process Model for Critical Data Analysis and Visualization. *arxiv.org*, 4 2022.

Emily Bhuwalka, Kunal; Icel, Nur; Gong. How does Robot Feedback Affect Participant Affinity and Trust ? *HRI*, 2018.

Scott D Blain, Matthew A Snodgress, Lauri Nummenmaa, Joel S Peterman, Enrico Glerean, and Sohee Park. Social bodies: Preliminary evidence that

awareness of embodied emotions is associated with recognition of emotions in the bodily cues of others. *Psychology of Consciousness: Theory, Research, and Practice*, 2023.

Kirsten Boehner, Rogério Roge´rio Depaula, Paul Dourish, and Phoebe 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`.

Christian Bors, Theresia Gschwandtner, and Silvia Miksch. Capturing and visualizing provenance from data wrangling. *IEEE computer graphics and applications*, 39(6):61–75, 2019.

Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

Virginia Braun and Victoria Clarke. *Thematic analysis.* American Psychological Association, 2012.

Virginia Braun and Victoria Clarke. Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, 9:3–26, 2 2022. doi: 10.1037/qup0000196. URL `https://doi.org/10.1037%2Fqup0000196`.

Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. Crowdsourcing Human-Robot Interaction: New Methods and System Evaluation in a Public Environment. *Journal of Human-Robot Interaction*, 2(1):82–111, 2013. doi: 10.5898/jhri.2.1.breazeal.

Harald Breivik, PC Borchgrevink, SM Allen, LA Rosseland, L Romundstad, EK Breivik Hals, G Kvarstein, and A Stubhaug. Assessment of pain. *BJA: British Journal of Anaesthesia*, 101(1):17–24, 2008.

Mason Bretan, Guy Hoffman, and Gil 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`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Paul Bucci, Xi Laura Cang, Anasazi Valair, David Marino, Lucia Tseng, Merel Jung, Jussi Rantala, Oliver S Schneider, and Karon E MacLean. Sketching cuddlebits: coupled prototyping of body and behaviour for an affective robot pet. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3681–3692. ACM, 2017.

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

Paul Bucci, Leo Foord-Kelcey, Patrick Yung Kang Lee, Alamjeet Singh, and Ivan Beschastnikh. Crystallizing Schemas with Teleoscope: Thematic Curation of Large Text Corpora, 2024. URL `https://arxiv.org/abs/2402.06124v2`.

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

David Byrne and Gillian Callaghan. *Complexity theory and the social sciences: The state of the art*. Routledge, 2013.

Walter B Cannon. The james-lange theory of emotions: a critical examination and an alternative theory. *The American journal of psychology*, 100 (3/4):567–586, 1987.

M Ariel Cascio, Rossio Motta-Ochoa, Gail Teachman, Florian Grond, Tamar Tembeck, Stephanie Blain-Moraes, and Melissa Park. What's at Stake with Biomusic? Ethical Reflections on an Emerging Technology. *Journal of Humanities in Rehabilitation*, 2020.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

ChatGPT. URL `https://openai.com/blog/chatgpt`.

Huili Chen, Hae Won Park, and Cynthia Breazeal. Teaching and learning with children: Impact of reciprocal peer learning with a social robot on

children's learning and emotive engagement. *Computers & Education*, 150:103836, 2020.

Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. Using Machine Learning to Support Qualitative Coding in Social Science. *ACM Transactions on Interactive Intelligent Systems*, 8: 1–20, 6 2018a. doi: 10.1145/3185515. URL `https://doi.org/10.1145%2F3185515`.

Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–20, 2018b.

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12):1992–2001, 2013.

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

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

Zach Cutler, Kiran Gadhave, and Alexander Lex. Trrack: A Library for Provenance-Tracking in Web-Based Visualizations. In *IEEE Visualization Conference (VIS)*, pages 116–120, 2020. doi: 10.1109/VIS47514.2020.00030.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*, 2023.

Oliver Damm, Karoline Dreier, Frank Hegel, Petra Jaecks, Prisca Stenneken, Britta Wrede, and Martina Hielscher-Fastabend. Communicating emotions in robotics: Towards a model of emotional alignment. In *Proceedings of the workshop" Expectations in intuitive interaction" on the 6th HRI International conference on Human-Robot Interaction*, 2011.

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

John Deigh. Cognitivism in the Theory of Emotions. *Ethics*, 104(4):824–854, 1994.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

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

Sam M Doesburg, Jessica J Green, John J McDonald, and Lawrence M Ward. Rhythms of consciousness: binocular rivalry reveals large-scale oscillatory network dynamics mediating visual perception. *PloS one*, 4 (7):e6142, 2009.

Panteleimon Ekkekakis. *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press, 2013.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200, 1992.

Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. Semantic concept spaces: Guided topic model refinement using word-embedding projections. *IEEE transactions on visualization and computer graphics*, 26(1):1001–1011, 2019.

Phoebe C Ellsworth. Appraisal theory: Old and new questions. *Emotion Review*, 5(2):125–131, 2013.

Anna Fariha and Alexandra Meliou. Example-driven query intent discovery: Abductive reasoning using semantic similarity. *arXiv preprint arXiv:1906.10322*, 2019.

Kerstin Fischer, Lars Christian Jensen, Maria Vanessa, and Der Wieschen. Emotion Expression in HRI – When and Why. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 29–38, 2019.

Joseph Fridman, Lisa Feldman Barrett, Jolie B Wormwood, and Karen S Quigley. Applying the theory of constructed emotion to police decision making. *Frontiers in psychology*, 10:1946, 2019.

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

Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, Tom Ewing, Keith N Hampton, and Naren Ramakrishnan. ThemeDelta: Dynamic segmentations over temporal topic models. *IEEE transactions on visualization and computer graphics*, 21(5):672–685, 2015.

Guido Gainotti. Neuropsychological theories of emotion. *The neuropsychology of emotion*, pages 214–236, 2000.

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

Ge Gao, Malte F. Jung, Gabriel Culbertson, Susan R Fussell, Malte F Jung, Sun Young Hwang, Gabriel Culbertson, Susan R Fussell, and Malte F Jung. Beyond Information Content: The Effects of Culture On Affective Grounding in Instant Messaging Conversations. *Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article*, 1(18):1–18, 2017. ISSN 2573-0142. doi: 10.1145/3134683. URL `https://doi.org/10.1145/3134683{%}0Ahttps://www.researchgate.net/publication/321193673`.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. CoAIcoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *ACM Transactions on Computer-Human Interaction*, 31(1):1–38, 2023.

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. CollabCoder: A Lower-barrier,

Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. *arXiv preprint*, 2024.

Greg Guest, Emily Namey, and Mario Chen. A simple method to assess and report thematic saturation in qualitative research. *PloS one*, 15(5): e0232076, 2020.

Marti A Hearst and Duane Degler. Sewing the seams of sensemaking: A practical interface for tagging and organizing saved search results. In *Proceedings of the symposium on human-computer interaction and information retrieval*, pages 1–10, 2013.

Monique M Hennink, Bonnie N Kaiser, and Vincent C Marconi. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research*, 27(4):591–608, 2017.

Gilad Hirschberger. Collective trauma and the social construction of meaning. *Frontiers in psychology*, 9:351992, 2018.

JR Hodges. Making it up and making do: Simulation, imagination, and empathic accuracy. *The Handbook of Imagination and Mental Simulation*, pages 281–294, 2008.

Tom Hollenstein. State space grids. In *State Space Grids*, pages 11–33. Springer, 2013.

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

Matt-Heun Hong, Lauren A Marsh, Jessica L Feuston, Janet Ruppert, Jed R Brubaker, and Danielle Albers Szafir. Scholastic: Graphical Human-AI Collaboration for Inductive and Interpretive Text Analysis. ACM, 10 2022. doi: 10.1145/3526113.3545681. URL `https://doi.org/10.1145%2F3526113.3545681`.

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.

Robert Isaacson. *The limbic system*. Springer Science & Business Media, 2013.

Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

Malte Jung and Pamela Hinds. Robots in the wild: A time for more robust theories of human-robot interaction, 2018.

Malte F. Jung. Affective Grounding in Human-Robot Interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pages 263–273, 2017. ISBN 9781450343367. doi: 10.1145/2909824.3020224. URL `http://dl.acm.org/citation.cfm?doid=2909824.3020224`.

Hannah Kim, Dongjin Choi, Barry Drake, Alex Endert, and Haesun Park. TopicSifter: Interactive search space reduction through targeted topic modeling. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 35–45, Vancouver, Canada, 2019. IEEE, IEEE.

Hannah Kim, Barry Drake, Alex Endert, and Haesun Park. Architext: Interactive hierarchical topic modeling. *IEEE transactions on visualization and computer graphics*, 27(9):3644–3655, 2020.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. MEGAnno+: A Human-LLM Collaborative Annotation System, 2024.

Bryan Kolb and Laughlin Taylor. Neocortical substrates of emotional behavior. In *Psychological and biological approaches to emotion*, pages 133–162. Psychology Press, 2013.

Jürgen Kornmeier and Michael Bach. The Necker cube—an ambiguous figure disambiguated in early visual processing. *Vision research*, 45(8): 955–960, 2005.

James H Kryklywy, Mana R Ehlers, Andre O Beukers, Sarah R Moore, Rebecca M Todd, and Adam K Anderson. Decomposing neural representational patterns of discriminatory and hedonic information during somatosensory stimulation. *eneuro*, 10(1), 2023.

Kori A LaDonna, Anthony R Artino Jr, and Dorene F Balmer. Beyond the guise of saturation: rigor and qualitative interview data, 2021.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

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

Peter J Lang. The varieties of emotional experience: a meditation on James-Lange theory. *Psychological review*, 101(2):211, 1994.

Lucian Leahu and Phoebe 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`.

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

Ching-Hung Lee, Chien-Liang Liu, Amy JC Trappey, John PT Mo, and Kevin C Desouza. Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery. *Advanced Engineering Informatics*, 50:101428, 2021.

Yuan Li, Anita Crescenzi, Austin R Ward, and Rob Capra. Thinking inside the box: An evaluation of a novel search-assisting tool for supporting (meta) cognition during exploratory search. *Journal of the Association for Information Science and Technology*, 2023.

Marsha Linehan. *DBT? Skills training manual*. Guilford Publications, 2014.

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

Matteo Lissandrini, Davide Mottin, Themis Palpanas, Yannis Velegrakis, and HV Jagadish. *Data Exploration Using Example-Based Methods*. Springer, 2019.

Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. *arXiv preprint arXiv:2310.02161*, 2023.

Russell A Lockhart. Interrelations between amplitude, latency, rise time, and the Edelberg recovery measure of the galvanic skin response. *Psychophysiology*, 9(4):437–442, 1972.

Stephen Madigan. *Narrative therapy.* American Psychological Association, 2011.

Kirsti Malterud, Volkert Dirk Siersma, and Ann Dorrit Guassora. Sample size in qualitative interview studies: guided by information power. *Qualitative health research*, 26(13):1753–1760, 2016.

Jean Matter Mandler. *Stories, scripts, and scenes: Aspects of schema theory.* Psychology Press, 2014.

Sara Mannheimer. Data curation implications of qualitative data reuse and big social research. *Journal of eScience Librarianship*, 10(4), 2021.

Renee CB Manworren and Jennifer Stinson. Pediatric pain measurement, assessment, and evaluation. In *Seminars in pediatric neurology*, volume 23.3, pages 189–200. Elsevier, 2016.

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

Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1), 2018. ISSN 20452322. doi: 10.1038/s41598-018-32063-4.

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

Denis Mayr Lima Martins. Reverse engineering database queries from examples: State-of-the-art, challenges, and research opportunities. *Information Systems*, 83:89–100, 2019.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, 3(29):861, 2018.

Christofer Meinecke, David Joseph Wrisley, and Stefan Jänicke. Explaining semi-supervised text alignment through visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4797–4809, 2021.

Derek L Milne and Robert P Reiser. *A manual for evidence-based CBT supervision*. John Wiley & Sons, 2017.

Adam S Miner, Sheridan A Stewart, Meghan C Halley, Laura K Nelson, and Eleni Linos. Formally comparing topic models and human-generated qualitative coding of physician mothers' experiences of workplace discrimination. *Big Data & Society*, 10(1):20539517221149106, 2023.

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

Albine Moser and Irene Korstjens. Series: Practical guidance to qualitative research. Part 1: Introduction. *European Journal of General Practice*, 23:271–273, 10 2017. doi: 10.1080/13814788.2017.1375093. URL `https://doi.org/10.1080%2F13814788.2017.1375093`.

Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

Toru Nakata, Tomomasa Sato, and Taketoshi Mori. Expression of emotion and intention by robot body movement. *5th Conference on Intelligent Autonomous Systems*, pages 352 – 359, 1998. 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`.

Jakob Neilson. 10 usability heuristics for user interface design. URL `https://www.nngroup.com/articles/ten-usability-heuristics/`.

Alan Michael Nevill and Andrew Lane. Why self-report "likert" scale data should not be log-transformed. *Journal of Sports Sciences*, 25(1):1–2, 2007. doi: 10.1080/02640410601111183. URL `https://doi.org/10.1080/02640410601111183`. PMID: 17127576.

Jakob Nielsen. Finding Usability Problems through Heuristic Evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 373–380, 1992. doi: 10.1145/142750.142834. URL `https://doi.org/10.1145/142750.142834`.

Jakob Nielsen and Rolf Molich. Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 249–256, 1990. doi: 10.1145/97243.97281. URL `https://doi.org/10.1145/97243.97281`.

Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102, 2017.

Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, 16(1):1609406917733847, 2017.

Lauri Nummenmaa and Riitta Hari. Bodily feelings and aesthetic experience of art. *Cognition and Emotion*, 37(3):515–528, 2023.

Juhani Ojala, Juulia T Suvilehto, Lauri Nummenmaa, and Eija Kalso. Bodily maps of emotions and pain: tactile and hedonic sensitivity in healthy controls and patients experiencing chronic pain. *Pain*, pages 10–1097, 2022.

KS Ong and RA Seymour. Pain measurement in humans. *The Surgeon*, 2 (1):15–27, 2004.

OpenAI. GPT-4 Technical Report, 2023.

Pablo E Paredes, Francisco Ordonez, Wendy Ju, and James A Landay. Fast & furious: detecting stress with a car steering wheel. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

André Parent and Lili-Naz Hazrati. Functional anatomy of the basal ganglia. i. the cortico-basal ganglia-thalamo-cortical loop. *Brain research reviews*, 20(1):91–127, 1995.

Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. A model-free affective reinforcement learning approach to

personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33.01, pages 687–694, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Michaela Pfadenhaueris and Hubert Knoblauch. Social constructivism as paradigm. *The legacy of the social construction of reality*, 2019.

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

Robert Plutchik. A psychoevolutionary theory of emotions, 1982.

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

Hilary Putnam. The meaning of "meaning". *Mind, Language, and Reality*, 1975.

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

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

Reddit.com. Am I the Asshole? `https://www.reddit.com/r/AmItheAsshole/`, 2024. Accessed: 2024-05-13.

Radim Rehurek and Petr Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

Tim Rietz and Alexander Maedche. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. ACM, 5 2021. doi: 10.1145/3411764.3445591. URL https://doi.org/10.1145%2F3411764.3445591.

Mark Risjord. *Nursing knowledge: Science, practice, and philosophy*. John Wiley & Sons, 2011.

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

Matthias Rüdiger, David Antons, Amol M Joshi, and Torsten-Oliver Salge. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *Plos one*, 17(4):e0266325, 2022.

James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

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

James A Russell. The greater constructionist project for emotion. *The psychological construction of emotion*, pages 429–447, 2015.

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

Martin Saerbeck and Christoph Bartneck. Perception of affect elicited by robot motion. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10*, page 53, 2010. ISBN 9781424448937. doi: 10.1145/1734454.1734473. URL https://www.bartneck.de/publications/2010/perceptionAffectElicitedRobotMotion/saerbeckBartneckHRI2010.pdfhttp://portal.acm.org/citation.cfm?doid=1734454.1734473.

Jelle Saldien, Kristof Goris, Bram Vanderborght, Johan Vanderfaeillie, and Dirk Lefeber. Expressing emotions with the social robot probo. *International Journal of Social Robotics*, 2(4):377–389, 2010.

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

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

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

Alexander Schmemann. *For the life of the world: sacraments and orthodoxy*. St Vladimir's Seminary Press, 1973.

Michael Schwarz. Is psychology based on a methodological error? *Integrative psychological and behavioral science*, 43(3):185–213, 2009.

John R Searle. *Seeing things as they are: A theory of perception*. Oxford University Press, 2015.

Favourate Y Sebele-Mpofu. Saturation controversy in qualitative research: Complexities and underlying assumptions. A literature review. *Cogent Social Sciences*, 6(1):1838706, 2020.

Solace Shen, Petr Slovak, and Malte F Jung. Stop. i see a conflict happening.: A robot mediator for young children's interpersonal conflict resolution. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 69–77. ACM, 2018.

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

Ali I Siam, Samah A Gamel, and Fatma M Talaat. Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques. *Neural Computing and Applications*, 35(17): 12891–12904, 2023.

Claudio T Silva, Juliana Freire, and Steven P Callahan. Provenance for visualizations: Reproducibility and beyond. *Computing in Science & Engineering*, 9(5):82–89, 2007.

Jeffry A Simpson. Foundations of interpersonal trust. *Social psychology: Handbook of basic principles*, 2:587–607, 2007.

Sichao Song and Seiji Yamada. Expressing emotions through color, sound, and vibration with an appearance-constrained social robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 2–11. ACM, 2017.

Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, D Horng Chau, Alex Endert, and Daniel Keim. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. In *Computer Graphics Forum*, volume 40.3, pages 543–568. Wiley Online Library, 2021.

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

Jennifer N Stinson. Improving the assessment of pediatric chronic pain: harnessing the potential of electronic diaries. *Pain Research and Management*, 14(1):59–64, 2009.

Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian 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.

Gail M Sullivan and Anthony R Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013.

Oleksandra Sushchenko, Enrico Glerean, Helena Sederholm, Lauri Nummenmaa, and Riitta Hari. Bodily Feelings and Emotional Landscapes of Aesthetic Experience: Overview of an Ongoing Research. In *End Games and Emotions: The Sense of Ending in Modern in Literature and Arts*, 2017.

Teleoscope.ca. Teleoscope. `https://teleoscope.ca`, 2024. Accessed: 2024-05-13.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. OCTIS: Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.31. URL `https://aclanthology.org/2021.eacl-demos.31`.

TICIRC. What is trauma-informed care? - trauma-informed care implementation resource center, Apr 2020. URL `https://www.traumainformedcare.chcs.org/what-is-trauma-informed-care/`.

Gerard A Tobin and Cecily M Begley. Methodological rigour within a qualitative framework. *Journal of advanced nursing*, 48(4):388–396, 2004.

Michael Tomasello. Joint attention as social cognition. In *Joint attention*, pages 103–130. Psychology Press, 2014.

Lénie J Torregrossa, Matthew A Snodgress, Seok Jin Hong, Heathman S Nichols, Enrico Glerean, Lauri Nummenmaa, and Sohee Park. Anomalous bodily maps of emotions in schizophrenia. *Schizophrenia bulletin*, 45(5): 1060–1067, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jonathan H Turner and Jan E Stets. Sociological theories of human emotions. *Annu. Rev. Sociol.*, 32:25–52, 2006.

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

Bessel A Van der Kolk. *The body keeps the score: Brain, mind, and body in the healing of trauma*. Penguin Books, 2015.

Janne Van Kollenburg, Sander Bogers, Heleen Rutjes, Eva Deckers, Joep Frens, and Caroline Hummels. Exploring the value of parent tracked baby data in interactions with healthcare professionals: A data-enabled design exploration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

Sofia Volynets, Enrico Glerean, Jari K Hietanen, Riitta Hari, and Lauri Nummenmaa. Bodily maps of emotions are culturally universal. *Emotion*, 20(7):1127, 2020.

Lawrence M Ward. Neuronal synchronization, attention orienting, and primary consciousness. *Multimodal Oscillation-based connectivity theory*, pages 29–49, 2016.

Lawrence M Ward, Sam M Doesburg, Keiichi Kitajo, Shannon E MacLean, and Alexa B Roggeveen. Neural synchrony in stochastic resonance, attention, and consciousness. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 60(4):319, 2006.

David Watson, Lee A Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988a.

David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988b.

David West. Language, thought and reality: a comparison of Ferdinand de Saussure's Course in General Linguistics with CK Ogden and IA Richards' The Meaning of Meaning. *Changing English*, 12(2):327–336, 2005.

Kory Westlund, M Jacqueline, Sooyeon Jeong, Hae W Park, Samuel Ronfard, Aradhana Adhikari, Paul L Harris, David DeSteno, and Cynthia L Breazeal. Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative. *Frontiers in human neuroscience*, 11:295, 2017.

Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. Thank you for sharing that interesting fact!: Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 298–306. ACM, 2018.

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

Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum*, volume 39.3, pages 757–783. Wiley Online Library, 2020.

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

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.

Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7:3–36, 2021.

Pamela Zontone, Antonio Affanni, Riccardo Bernardini, Alessandro Piras, and Roberto Rinaldo. Stress detection through electrodermal activity (EDA) and electrocardiogram (ECG) analysis in car drivers. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

Appendix A

Papers Referenced for Meta-Analysis

- Cang, X. L., Guerra, R. R., Guta, B., Bucci, P., Rodgers, L., Mah, H., ... & MacLean, K. E. (2023). FEELing (key) Pressed: Implicit Touch Pressure Bests Brain Activity in Modelling Emotion Dynamics in the Space Between Stressed and Relaxed. IEEE Transactions on Haptics.
- Cang, X. L., Guerra, R. R., Bucci, P., Guta, B., MacLean, K., Rodgers, L., ... & Agrawal, A. (2022, October). Choose or fuse: Enriching data views with multi-label emotion dynamics. In 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-8). IEEE.
- Cang, X. L., Bucci, P., Rantala, J., & MacLean, K. E. (2021). Discerning affect from touch and gaze during interaction with a robot pet. IEEE Transactions on Affective Computing, 14(2), 1598-1612.
- Bucci, P., Zhang, L., Cang, X. L., & MacLean, K. E. (2018, April). 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 (pp. 1-11).
- Marino, D., Bucci, P., Schneider, O. S., & MacLean, K. E. (2017, June). Voodle: Vocal doodling to sketch affective robot motion. In Proceedings of the 2017 Conference on Designing Interactive Systems (pp. 753-765).
- Bucci, P., Cang, X. L., Valair, A., Marino, D., Tseng, L., Jung, M., ... & MacLean, K. E. (2017, May). 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 (pp. 3681-3692).

- Cang, X. L., Bucci, P., Strang, A., Allen, J., MacLean, K., & Liu, H. S. (2015, November). Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 147-154).