Essays on Mind-Reading Machines

by

Nick Merrill

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

School of Information

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Chuang, Chair
Associate Professor Coye Cheshire
Associate Professor Alva Noë

Spring 2018

Essays on Mind-Reading Machines

by

Nick Merrill

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

School of Information

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Chuang, Chair
Associate Professor Coye Cheshire
Associate Professor Alva Noë

Spring 2018

The dissertation of Nick Merrill, titled Essays on Mind-Reading Machines, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Essays on Mind-Reading Machines

Copyright 2018
by
Nick Merrill

Abstract

Essays on Mind-Reading Machines

by

Nick Merrill

Doctor of Philosophy in School of Information

University of California, Berkeley

Professor John Chuang, Chair

Invasive brag; forbearance.

To Ossie Bernosky

And exposition? Of go. No upstairs do fingering. Or obstructive, or purposeful. In the glitter. For so talented. Which is confines cocoa accomplished. Masterpiece as devoted. My primal the narcotic. For cine? To by recollection bleeding. That calf are infant. In clause. Be a popularly. A as midnight transcript alike. Washable an acre. To canned, silence in foreign.

Contents

List of Figures

List of Tables

Acknowledgments

I want to thank my advisor for advising me.

Chapter 1

Mind-reading & telepathy for beginners and intermediates

This chapter outlines how theories of a mind that is embodied (beyond the brain), distributed (beyond the individual), and extended (to the built environment) make the mind amenable to sensing by emerging classes of wearable and environmental sensors.

Would you wear a device in the workplace if your manager used it to track your productivity, or creativity? Would you allow your child wear the same device in schools, where it could monitor both their academic achievement and their mental health? Would you wear a fitness tracker if your resting heartrate could predict your future involvement in violent crime?

In all of these examples, sensing technologies blur the line between *sensing bodies* and *sensing minds*. Today, increasingly inexpensive sensors with developer-friendly SDKs and APIs allow those with requisite software expertise to purport to detect phenomena ranging from focus to mental health to mood—all without direct data about the brain.

In this paper, I seek to dethrone the assumption that brain-scanning is necessary for computers to "read" or "decode" the mind. Drawing from contemporary theories of embodied, extended and distributed cognition, I argue that pervasive biosensors are already able to grasp at the contents of our minds by sensing our bodies, tools and built environment (Section 1.1). I relate this argument to ongoing work affective computing, computational social science, and biosensing, reframing these research programs as having already begun the work of theorizing and building computational *models of minds* (Section 1.2).

Drawing on critiques of this past work, I center the primacy of human interpretation in both constructing MoMs and interpreting their relevance in daily life. I propose these processes of construction and interpretation as a starting point for understand what MoMs are, and how they might operate in the world (Section 1.3).

1.1 Background

Consider the ant. The fungal complex *Ophiocordyceps unilateralis sensu lato* overtakes its behavior without operating on its brain at all (Figure 1.1). Instead, it uses the ant's body to feed itself, constructing a network of coordinated sensing and actuation atop the ant's muscles [**Fredericksen2017**].

Figure 1.1: *Ophiocordyceps unilateralis sensu lato* takes control of an ant's mind without input from its brain. It forces the ant to chews on the underside of a twig, after which the ant's body will serve only as a medium for fungal reproduction.

Ignoring questions of control, consider the degree of *sensing* the fungus must perform in order to maintain and use the ant's body. Using the ant's existing bodily infrastructure, the fungus creates a *model* of ant-experience robust enough to control the organism completely. Although the *Ophiocordyceps* fungal complex cannot read the ant's brain (it has no physical presence there), it can read the ant's *mind* well enough to model its environment and body. The fungus's model of ant-experience may not be the same, or even similar, to those used by the host ant. Regardless, they are of a sufficient resolution to allow the fungus to acheive its (reproductive) goals.

With this fungus in mind, consider the emerging class of IoT devices, which are increasingly embedded in the built environment, worn on the body, worn *in* the body via ingestible pills. Though mundane, cameras too sense bodies often in public, often without subjects' knowledge [**DBLP:journals/corr/SedenbergWC17**]. In all of these examples, connected devices are endowed to some degree with the capacity to sense (that is, to build models of

[**Gitelman**]) human bodies in space. Past work refers to this process broadly as *biosensing*, and these devices as *biosensors* [**day2016biosensing**].

While humans are significantly more complex than ants, this fungus helps illustrate the possibility of creating *models of minds* with limited or no information from the brain. As I review in this section, contemporary philosophical theories engage seriously with the notion of a *beyond-the-brain mind*. I continue in the following section to argue that these theories (unwittingly) provide grounds to claim that even current biosensors can build *models of minds*.

Contemporary (material) theories of mind

What is the mind? What is its relationship to the body, and to the physical world? Philosophers have proposed two basic catagories for answers to this question. *Dualism* posits that the mind has non-physical components, whereas *physicalism* posits a mind of only physical components. (For a slice of this debate, see [**Chalmers1998**]).

The physicalist stance of mind squares with Jane Bennett's account of *thing-materialism*, in which things in the world have an intrinsic power, locked in networks of interactions with other things [**Bennett2013a**]. In turn, this materialism squares the physicalist account of mind with the project of biosensing [**day2016biosensing**, **nafus2016quantified**]. If mental phenomena are physical, then mental phenomena are potentially the subject of sensing. Bennett's *Thing-Power* at once justifies the strict physicalism that allows biosensing to refer to phenomena in the mind, and justifies the systems of reference that make observations from biosensors *about* the mind.

Cognitive science

Moreover, the physicalist interpretation lends itself naturally to scientific study. From the physicalist perspective, all phenomena in the mind can, be reduced to descriptions of physical activity, some physical theory could eventually explain the mind in entirety. Cognitive science has historically been an influential source of physicalist theories about the mind. It takes a computational account of the brain, understanding how it "processes information" [**Winograd1987a**] within the physical constraints of computational space and time. Its questions operate above the biological concerns of neuroscience, but below the behavioral concerns of psychology, offering computational *models* of "cognition."

As physical models can inform structural engineering, cognitive scientific models have informed research across psychology, artificial intelligence, and design [**Agre1997**]. Some early studies provided evidence for computational complexity in time-related tasks [**shepard1971mental**], which have in turned inspired numerous models of cognition. These models informed the design of neural networks, before the relatively recent discovery of performant backpropogation algorithms made neural networks practical to deploy [**minsky1969perceptrons**].

However, cognitive science has received considerable criticism [**Noe2004**, **Winograd1987a**]. Two critiques relevant to this paper focus on cognitive science's "isolationist assumtions": a

focus on the brain (isolated from the body), and a focus on the individual (isolated from social context, and from the environment). The following three sections review major responses to these critiques: embodied cognition, distributed cognition, and extended cognition.

Embodied cognition

Cognitive science's isolation of the brain rests on the belief that the brain is strictly equivalent to the mind. This assumption has encountered two primary critiques. First, the dichotomy between the brain and body is intrinsically unstable; neurons occur body-wide, running directly to the brain, such that it is difficult to evaluate the role of cerebral neural activity in the functions of mind irrespective of non-cereberal neural activity. Second, to quote Noë and Thompson (2004), "The exact way organisms are embodied simultaneously constrains and prescribes certain interactions within the environment." [**Noe2004**]. In other words, aspects of mind exist because of the physical conditions of an organism's body in the world.

These critiques ultimately resulted in the *Embodiement thesis*: that an agent's beyond-the-brain body plays a causal role in that agent's cognitive processing. As one example, Noë and O'Regan's analysis of vision recasts the "visual processing" of cognitive science, in which internal representations are built and manipulated within the brain, to an active, embodied process, in which the seen world provides its own representations, which the body and brain must meet through an active process of co-adaptation [**ORegan2001a**]. In its account of vision as an active process of co-construction, this analysis shares with some work in feminist epistemology, e.g. [**Haraway1988b**].

Extended and distributed cognition

While embodied cognition addresses the critique of cognitive science's isolation of the brain, it does not cognitive models' isolation of the indivdual from social or environmental context. Further, while the embodiement thesis points out the causal relationship between mind and the physical conditions of the body, it glosses over the relationship between these bodies and the world in which they are situated. In response, Clark's Extended Cognition thesis argues that the environment at large can be considered as part of the mind. Clark et al propose that "technological resources such as pens, paper, and personal computers are now so deeply integrated into our everyday lives that we couldn't accomplish many of our cognitive goals and purposes without them" [**Clark1998**].

This theory does not stop at tools, however, in describing a mind beyond the body. Broadly, extended cognition refocuses the brain away from the individual body, and toward the general "active role of the environment in shaping cognition" [**Clark1998**]. This theory paved the way towad a socially-extended cognition, or "distributed cognition," as proposed in Hutchins' (1995) ethnography of sailors on a naval vessel [**hutchins1995cognition**]. In his analysis, multiple individuals and the material environment play constituent roles in cognition, a mind that is distributed across multiple human and non-human actors.

The theories in this section establish make various cases for a mind that extends beyond the confines of the brain, and even beyond the confines of the body. With these theories in mind, the next section argues these theories make the mind amenable to modeling via sensors that are worn or embedded in the environment.

1.2 Models of minds

TODO intro

The theories outlined in the previous section all propose that the mind is physically instantiated in the material world. They differ only on *where* this mind exists, and where it does the work of cognition. Embodied cognition focuses on the body's role in cognition, where extended and distributed cognition theorize cognition as a process distributed across human and non-human actors.

Using embodied distributed, and extended cognition, this section argues that contemporary and emerging biosensors are already able to sense the human mind through interactions with the body and built environment, and through the constituency of sensing devices in cognition and experience. I use two strands of existing work, reading each through a different account of mind: affective computing through the lens of embodied cognition, and computational social science through the lens of distributed and embodied cognition. Through these readings, I recast each program of pervasive sensing as a program of *reading the mind*.

To assist in this analysis, I propose the notion of *models of minds*. This term borrows from the term *theory of mind* which, in autism research, refers to the ability to reason about mental states [**Baron-Cohen1995**]. By substituting the word "theory" with the word "model," I emphasize the notion of formal or algorithmic representations. By then turning this "model of mind" (singular) into *models of minds* (plural), I center the intrinsic contestibility of the algorithms that build them, and the beliefs that underlie their construction, as well as the many minds in the world to model. The term aims to cast a subtle doubt on models that appear too simple, or which claim to generalize across all people.

Building *models of minds* can be split into two major components: the designerly program of building algorithms that encode and represent mental states, and the social processes of understanding these representations as relevant in the course of life. While the boundary between these components is intrinsically unstable, this agential split is nonetheless useful in understanding how these models perform work in the world. To these two components, I assign the terms *telepathy* and *mind-reading*, respectively. While these two terms, especially the latter, have a strong connection to magic, I attempt briefly to explain the useufulness of repurposing them for discussions on computational models of minds. Consider telepathy's etymological pedigree in relation to other popular technologies.

Telephony (*tele* + *phonos*)
: Sound at a distance

Television (*tele + vīsiō*)
Sight at a distance

Telepathy (*tele + pathos*)
Mind at a distance.

While the first two terms may have sounded like magic at some point in history, technical infrastructures have provided functionality that made these terms legible not just as technologies but as social media. *Telepathy* is in spirit no different. In relation to the other technical infrastructures, the prefix *tele-* highlights technical aspects of transmission, along with the various sociotechnical infrastructures and entanglements that make transmission, encoding, and decoding both possible and desireable.

In contrast, *mind-reading*, a term associated with street and stage magic [**Ali2014a**], re-focuses on human aspects of interpretability and legibility. It centers the processes of meaning-making and the performance of understanding. Together, *mind-reading* and *telepathy* work to describe how models of minds are "made and measured" [**Boehner2007b**], while gesturing toward the unstable boundary between these two activities.

The remainder of this section reads work in affective computing through the lens of embodied cognition, and computational social science thruogh the lens of distributed and extended cognition. Through these readings, I demonstrate how existing work in biosensing has already built models of minds, and in so doing, offer various examples of what mind-reading and telepathy might mean in practice. These analyses provides context for a discussion in the following section on the primacy of human beliefs—both on the part of designers and users—in structuring mind-reading and telepathy.

Affective computing

Affective computing, pioneered by Rosalind Picard at the MIT Media Lab, sought to QUOTE PICARD+CITE. Affective computing seeks to model emotions in order to improve user interfaces with machines. A central vision of this project is to endow computers with the ability to perform emotions. Tightly intertwined with this goal, affective computing seeks to detect emotions in users. In this program, affective computing is concerned with mind-reading in the sense that other people perform it: the construction of a "theory of mind" [**Baron-Cohen1995**].

Affective computing, as it has been framed historically, has received a variety of critiques. First, work in affective computing has tended to frame emotions as definite entities in the world, for which a ground truth exists. Boehner et al [**Boehner2007b**] posit an alternative interpretation of emotions as co-constructed, performed socially and understood only in collaboration with other socially-experiencing subjects. An account of socially-situated

emotions has received some limited uptake within affective computing [**Parkinson2015**]; however, these theoreis still pre-catagorize emotions, which may obscure other phenomena at the borders of these categories [**Boehner2007b**].

Second, affective computing has not substantively engaged with the question of how algorithms and devices that seek to detect emotion may affect the way emotion is experienced or performed. To quote Cecil Adams, "the act of observing disturbs the observed" [**Adams1982**]. Feedback about emotional experience may alter emotional experience as well [CITE].

well with all these critiques, why are we bothering to talk about affective computing? _because it relies on notions of emotion as embodied states that can be detected from physiological input [HEALEY] _relies on data from the body to encode and transmit emotion - an aspect of mind.

TODO Computational social science

Past work has attempted to perform social science using sensors worn or embedded in the build environment. These efforts predate contemporary IoT, existing mostly in labs. One early example is Sandy Pentland's Sociometer, which I can describe a little. ubicomp, proximate future.

Describe Social fMRI as a seminal example. Explain how computational social science is **actually** telepathy — implicitly uses distributed cognition to understand mental phenomena like stress, anxiety, depression etc. Aided by infrastructures of machine learning that require large, multi-subject corpora, finding relational and longitudinal dependencies in the dataset.

Indeed, the world of computational social science has shifted from proximate future to lived present. From the commerical world of target advertisements, to the public-sector world of pervasive surveillance. The legacy of Pentland's work lives on in our connected present. easy critique that looked from perspective of manager, sought to make workspaces more efficient etc This top-down perview of the scientist ignored potential concerns around individual privacy, a legacy that continues to produce struggles in IoT [god view]. Perfect place for a transition to conclusion.

TRAJECTORIES OF DEPRESSION if u believe depression to be an embodied phenomenon, then the phone does sense depression through the body, reather than its bodily correlates. if u believe depression to be an extended phenomenon, then the cellphone is in fact *a constituent part of the dep* and can report the ground truth of the depression.

these stances are relevant in understanding how mind-reading and telepathy are constructed via the interaction of human beliefs and material configurations, how these theories make it possible to seek ground truth such that models can be said to be (in)accurate. I discuss the centrality of human intepretation in the following section.

1.3 Centrality of interpretation

How does one read the mind? In short, one comes up with a theory in which the mind is made amenable to sensing. In this section, DESCRIBE PROJECT OF UDNERSTANDING HOW ENGINEERS AND RESEARCHERS CONSTRUCT MODELS OF MINDS, AND HOW USERS INTERPRET THEM.

TODO cite boehner first of all

User interpretations

Construction of engineers

Writing that exists somehow

If people *think* a certain technology measures aspects of mind, it will certainly affect the way they engage with that technology—whether or not it works the way they expect [**Ali2014a**]. Meanwhile, if they think that a given technology does **not** measure their mind, when it fact it does, users may suffer a breach of what Nissenbaum might call the "appropriateness of the flow of information" [**Doyle2011**]. In both cases, knowing what people expected will help us anticipate their needs, and concerns.

There are some people who actually *want* their minds measured. In these cases, technologies that claim to "measure the mind" must rely on end-users to define the criteria by which systems are deemed effective, or accurate. Consider the Spire, a breath sensor that claims to divine, from a person's patterns of in-breaths and out-breaths, what the user is calm, focused, or tense [**SpireInc**]. For the device to "work," not only must these detected signals match with end-users' intuitions, but users must also believe that a device like the Spire has the power to measure and detect these phenomena, given breath as input [**Ali2014a**]. Of course, these attitudes will not be fixed: they will evolve over time, as users observe the device in action, and correlate its judgments with their own lived experiences [**Nafus2016**].

Crucially, assumed connections between the readings of these sensors and the phenomena they purport to measure can break down. A cat could accidentally flip a lightswitch [**Tolmie2016**]; a heartrate sensor could give a false reading. Mismatches between what a signal is "supposed to" refer to, and what it actually refers to, highlight the other crucial aspect of biosensing: *that it relies on beliefs about the physical world*, baseline hypotheses that allow readings from these sensors to be classified as accurate or inaccurate [**Boyd2013**]. And, of course, "raw data is an oxymoron" [**Gitelman**]; the way data are presented in specific contexts define what human interpreters are and are not able to see [**Day2014**]. One useful point of comparison can be found in Nafus's (2016) [**Nafus2016**] description of how she began her studies of biosensing.

> Figuring out whether a consumer market for biosensors was even thinkable had everything to do with whether the data they produced cohered with a cultural and social imaginary, such that users stood a chance of making sense of them.

This revelation belies the legibility that allows biosensing to become meaningful in daily life. it is the step of active, interpretive construction that turns sensing into biosensing, as it takes on reference to living systems, or a MoM, as it takes reference to a human mind.

In other words, the way the mind is conceived has consequences for the way technoogies are employed to sense it [**Rose2016**]. Only someone who believes emotion is socially situated would sense an entire organization to understand how someone is feeling [**OlguinOlguin2009**, **Healey2010**, **Aharony2011**]. Past work has touched on how preexisting categories of phenomenal experience *make* these phenomena amenable to detection or classification [**Boehner2007b**]. Theories of how the mind is physically instantiated drive the pursuit of detecting particular phenomena. After all, bodies are not transparent. An inherent ambiguity exists in what data from sensors can mean, or what it would mean for these data to be accurate.

Describe project of knowing what end-users **and engineers** think aout this. Contrast with STS proiject of scientists in labs. Now, users and engineers outside of lab environments will parttake, wittingly or not, and with varying degrees of formality, in this centuries-old project of Western rationalist inquiry. Lead into the next section.

Chapter 2

Reading mind from heartrate

While the previous chapter argues that certain framings of the mind make it amenable to reading by sensors, this chapter seeks to discover whether *users believe* that biosensors can capture aspects of mind. Through a vignette experiment and a mixed-methods experimental study, this chapter show how people use biosensory data (heartrate) in social, computer-mediated contexts to build interpretations relating to the minds of others.

2.1 Introduction

As of 2016, several apps allow users to share their heartrate with their friends, leading some [**McNell2015**] to wonder why anyone would anyone want to do such a thing. In fact, heartrate is a potentially rich signal for designers. The meaning of a heartrate in any given context is at once socially informative [**Frey2016a**, **Slovak2012**] and highly ambiguous [**Merrill2010a**].

After all, heartrate is not just some number. The sense of one's heartbeat is an integral feature of the human experience, and people's associations with it range from intimacy [**Janssen2010**] to anxiety [**Decaria1974**] to sexual arousal [**Valins1966**]. Many heartrate sharing applications rely on these associations, asking users to ascribe contextual meanings to heartrate [**Kastrenakes2014**, **Slovak2012**], often with the aim of increasing intimacy [**Janssen2010**]. The advertising copy for Cardiogr.am, one smartwatch app, reads,

> Your heart beats 102,000 times per day, and it reacts to everything that happens in your life—what you're eating, how you exercise, a stressful moment, or a happy memory. What's your heart telling you?

These applications, along with many others, rely on the fact that people will imbue their heartrate data with emotional, and highly contextual interpretations. Given the relatively large number of wearables with embedded heartrate monitors (watches, bands, even earbuds) [**Stables2016**], it is unsurprising that designers are looking beyond fitness and health for

ways to increase user engagement with these devices. However, it is not clear how individuals will interpret a shared biosignal (e.g. heartrate) in different contexts of social interaction.

This chapter examines what heartrate can mean as a computer-mediated cue, and how interpretations of heartrate affect social attitudes and social behavior as people assign meanings to these signals relevant to the mind (emotion, mood, trust).

First, we use a vignette experiment to investigate how individuals make social interpretations about a rudimentary biosignal (heartrate) in conditions of uncertainty, focusing on dyadic interactions between acquaintances. Dyadic relations, which are present in all groups, function as a fundamental starting point for understanding interpersonal collaboration and group interactions [**Cheshire2010**]. We describe the quantitative and qualitative results of a randomized vignette experiment in which subjects make assessments about an acquaintance based on an imagined scenario that included shared heartrate information. We examine two contexts in this study: an uncertain, non-adversarial context and an uncertain, adversarial context. These two contexts, differing only by a few words, ask participants to imagine they are meeting someone "for a movie" (non-adversarial) or "to discuss a legal dispute" (adversarial).

We find that a high heartrate transmits negative cues about mood in both contexts of interaction, but that these cues do not appear to impact assessments of trustworthiness, reliability or dependability. Counter to our initial predictions, we find that normal (rather than elevated) heartrate leads to negative trust-related assessments, but only in the adversarial context. In qualitative assessments of subjects' attitudes and beliefs, we find that normal heartrate in the adversarial condition conflicts with expectations about how the participant believes the acquaintance should feel, signaling a lack of concern or seriousness, which appears to lead individuals to view the acquaintance as less trustworthy. In contrast, subjects in the non-adversarial context relate elevated heartrate to empathy and identification rather than trustworthiness. We also find a small number of subjects read different social interpretations onto the heartrate signal, including a very small minority who did not infer any relationship between the heartrate and the social situation.

From these findings about social attitudes, we then move to an lab-based experiment to understand how shared heartrate effects social behavior. We apply quantitative and qualitative analyses to an iterated prisoner's dilemma game, in which heartrate information ("elevated" or "normal") was shared between players. In a follow-up study, we replicate our initial study, but replace heartrate with an unfamiliar biosignal, "Skin Reflectivity Index (SRI)."

Our results raise important questions for applications that transmit sensor-derived signals socially between users. For signals with strong cultural associations, people's prior beliefs will color their interpretations, and social outcomes may or may not be positive. In the case of novel signals, on the other hand, our results imply that designers can (perhaps inadvertently) teach users to associate these biosignals with social meanings. This effect could be viewed as beneficial, depending on design objectives. It could also be dangerous if designers suggest, perhaps even inadvertently, interpretations that lead to discrimination.

2.2 Background

Sharing sensor data

To date, most work on the contextual interpretation of sensor data has focused on individual interpretation of individual data (c.f. quantified self). In contrast, our work attempts to move toward an understanding of how biosignals are interpreted in interpersonal interactions – the quantified social self. This shift is motivated, in part, by an increasing number of consumer applications that support sharing biosignals such as heartrate. Especially pertinent to our study, it is not well understood what heartrate actually signals to another person in a social interaction. How might the contextual, social interpretation of another person's biosignals affect social interpretations of mood (e.g., anxiety, calmness), or attitudes about trustworthiness and dependability?

Goffman [**Goffmann1959**] (p 56) makes an important distinction between the cues that we intend to give to others, and those that are "given off" unintentionally through our numerous non-verbal actions and behaviors. We view physiological signals such as heartrate as a form of non-verbal signaling that can "give off" more information to others than the sender may desire [**Howell2016**]. This type of personal data revealed through discreet sensors paired with mobile communication technologies has, until recently, been unavailable in most forms of social interaction.

Sharing physiological data

Prior work interrogates the contextual interpretation of personal data from certain kinds of sensors [**Choe2011a**, **Consolvo2005**], but physiological data has received less attention, despite two crucial differences from sensors that capture information such as location (e.g., GPS). First, biosensor data are intrinsically ambiguous: whereas a GPS coordinate refers to one specific place, heartrates do not have oneto-one mappings to physical activities or emotions. Second, physiological phenomena vary from person to person; 60bpm could be high or low depending on whose heartrate it is. A relatively large body of work has looked at how the transmission of physiological data might play a role in computermediated communication. One class of application has attempted to explicitly encourage or discourage certain behavioral outcomes, making some biosignals apparent such that the transmission of the data acts as a social cue [**Bergstrom2011**]. Another class of prototypes explores how signals might affect feelings of intimacy, particularly between romantic partners [**Bell2003**], and several applications focus on the transmission of heartrate as a means to achieve this effect [**Janssen2010**, **McNell2015**].

Sharing heartrate

Heartrate has deep-rooted cultural significance in many societies, and near-universal familiarity as a feature of our lived experiences. Building on associations with intimacy and

love, many heartrate sharing applications have aimed to "enhance" social connectedness by fostering feelings of intimacy [**Janssen2010**, **hassibheartchat**] between people.

What heartrate means as a computer-mediated cue, however, is ambiguous, its potential interpretations varying widely in different contexts [**Lotan2007**, **Slovak2012**]. Boehner et al (2007) argue for the intrinsic ambiguity of sensor data as a resource in design, particularly in systems that seek to use these data to express emotion [**Boehner2007b**]. Many technology probes corroborate this stance, relying on users to project socially contextual meanings around a transmitted heartrate. Consequently, more recent work has challenged the notion that the social consequences of transmitting physiological data will always result in increased trust and intimacy. There remains little work, however, on how the potential ambiguity of a heartrate signal is resolved in social conditions of risk and uncertainty.

2.3 Vignette experiment

This section describes the quantitative and qualitative results of a randomized vignette experiment in which subjects were asked to make assessments about an acquaintance based on an imagined scenario that included shared heartrate information. We compare the results of this experiment in adversarial and non-adversarial contexts of interaction. We find that elevated heartrate transmits cues about mood in both contexts, but that these cues do not appear to impact assessments of trustworthiness, reliability and dependability. Counter to our expectations, we find that normal (rather than elevated) heartrate leads to negative trust-related assessments, but only in an adversarial context. Our qualitative analysis points to the role of social expectations in shaping contextual interpretations of heartrate, and reveals individual differences in the way interpretations are constructed. We unpack some of the ways that social meanings can arise from biosensor data, and discuss considerations for those designing interactions with wearables.

Compared to social interpretations of physiological signals, interpretations of one's own signals are slightly better-understood from empirical research. Individuals' interpretations of their own heartrate have received particular attention (see [**Parkinson1985**] for a review). Studies have generally revealed that, when individuals believe that their heartrate is elevated, they sometimes believe their mood and emotions to be more negative [**Young1982a**].

If lay interpretations of one's own heartrate can yield negative self-interpretations, sharing heartrate information could also yield negative effects on mood and trustworthiness, particularly during uncertain interactions where something is at stake (such as time, money, or other valued resources). To investigate, we use a mixed-methods approach combining quantitative and qualitative analyses of a survey-based vignette experiment.

Hypotheses

Based on aforementioned studies of individual's negative emotional interpretation of their own heartrate, we believe that this negative valence will be mirrored in people's interpre-

tations of the heartrates of others in uncertain situations. Our investigation begins with two key predictions about negative assessments of one's partner in an uncertain social situation. We test both hypotheses in two different contexts of interaction (adversarial and non-adversarial) to understand how the context of risk and uncertainty affects social interpretations of heartrate.

1. Heartrate and Mood

   Past work indicates that people tend to make negative inferences about mood and emotion from elevated heartrates [**Decaria1974**, **Gu2012**, **Young1982**]. As such, our first hypothesis predicts that participants will adjust their attitudes about the mood of their partner when their partner's heartrate is elevated, as opposed to normal: Hypothesis 1: When individuals believe that their partner has an elevated heartrate in an uncertain social interaction, they will report their partner as being (1a), less calm (1b), more emotional (1c), and more easily upset (1d), compared to those who believe that their partner has a normal heartrate.

2. Heartrate and Trustworthiness

   Where Hypothesis 1 predicts that individuals will make negative assessments about an acquaintance's mood based on elevated heartrate, our second hypothesis predicts that individuals will make negative assessments about dispositions to behave in a reliable, dependable and trustworthy manner. Thus, both hypotheses stem from the same base assumption that, all things being equal, elevated heartrate has a primarily negative connotation with attitudes and behaviors of another person. Hypothesis 2: When individuals believe that their partner has an elevated heartrate in an uncertain social interaction, they will make negative assessments about the partner's trustworthiness (2a), reliability (2b), and dependability (2c), compared to those who believe that their partner has a normal heartrate.

Methods

To test our hypotheses, we conducted a survey-based vignette experiment. Vignette studies involve short descriptions of a scenario, designed to elucidate opinions, attitudes, and beliefs about that particular situation [**Jenkins2010**].

In this vignette study, we compare two different contexts of interaction. We do not create separate hypotheses for the two different contexts; rather, we are interested in comparing and contrasting the two different contexts to see how they might interact with social interpretations of heartrate. We provide our participants with either an adversarial or a non-adversarial social context. In the adversarial scenario, the participant is waiting to meet an acquaintance about a legal dispute. In the non-adversarial scenario, the participant is waiting at a movie theater for an acquaintance so that they can see a film together.

In all scenarios, the acquaintance sends a message via smartphone indicating that he or she is running late due to slow traffic. The person who is waiting does not know if the

acquaintance will make it on time or not, or whether the acquaintance is being honest about their tardiness. Within each context, we manipulate a small piece of information about the heartrate of the acquaintance: We tell the participant that the heartrate of the acquaintance has been shared by the acquaintances' smartphone and it is either elevated or normal.

1. Sample

   Our sample was undergraduate students recruited from the population of a large, public university on the West Coast of the U.S. Potential participants were asked to participate in a short online survey, and they did not know the nature of the questions or the topic of the study in advance. All participants were paid a $5 Amazon gift card. One hundred and three participants (103) completed the experiment survey instrument. The pool was weighted toward women; in our sample, 65% were women and 34% are male, and 2% (2 subjects) did not identify with either gender. With random assignment, the same overall gender split was maintained across conditions. The mean age of participants was 23.

2. Vignettes Each participant in the study saw only one of the four possible vignettes. After the vignette, the survey included free response questions about subjects' reactions to and interpretations of the situation described in the vignette, as well as 7-point Likert scale questions (Strongly Agree to Strongly Disagree) in which subjects evaluated the other person's disposition ("This person is emotional", "This person is anxious", "This person is easily upset", and "This person is calm"). In addition, we asked participants to indicate whether the other person was "trustworthy," "reliable," and "dependable" using the same 7- point agreement scale.

   There are two contexts of interaction (adversarial and nonadversarial) and two heartrate conditions (normal and elevated), creating four distinct vignettes based on social context and heartrate (HR): adversarial elevated HR, adversarial normal HR, non-adversarial elevated HR, and non-adversarial normal HR. Participants were randomly assigned into one of the four conditions. We manipulated these heartrate conditions by making a key wording change as indicated in the two context vignettes below.

   a) Non-Adversarial

      > You planned to meet your acquaintance for a movie at seven. It's 7:15, and you're standing alone in front of the theater. Your phone buzzes, and you see a message from this person that says, "I'm running late, traffic was really slow." Through your smartphone, you are able to see this person's heartrate, which the app designates as [normal / elevated]. It is currently 75 degrees and sunny. Your movie starts at 7:20.

   b) Adversarial:

      > You planned to meet your acquaintance at seven to discuss a difficult legal dispute between the two of you. It's 7:15, and you're standing alone

> in front of the meeting spot. Your phone buzzes, and you see a message from this person that says, "I'm running late, traffic was really slow." Through your smartphone, you are able to see this person's heartrate, which the app designates as [normal / elevated]. It is currently 75 degrees and sunny.

Quantitative results

We apply both quantitative and qualitative analyses to investigate our research questions and hypotheses. The study is based around an experimental design, but we also place significant emphasis on open-ended responses to better understand participants' thought processes, beliefs, and rationale for their choices in the vignettes. Our first hypothesis predicts that individuals will make negative attributions about the mood of the acquaintance in this uncertain situation when they believe that the acquaintance has an elevated heartrate (compared to normal heartrate). Given our four separate measures of mood, we conducted a multivariate analysis of variance (MANOVA) to test the hypothesis that there are one or more mean differences between the normal/elevated heartrate conditions, and/or between the two contexts of interaction (nonadversarial and adversarial).

We found a strong, statistically significant effect and a medium practical association between emotional attributions and heartrate condition, $F\ (4, 96) = 32.89$, $p < .001$; partial eta squared = .58. Turning to the individual outcomes, we find that subjects' perceptions of the acquaintance in the vignette's anxiety, his/her tendency to be easily upset, his/her tendency to be emotional, and his/her lack of calmness were all significantly higher in the elevated heartrate conditions when compared to the normal heartrate conditions (see Figure 2.1). We found no significant effect for the two contexts of interaction, $F\ (4, 96) = 1.072$, $p = .38$, and no significant effect for the context x heartrate condition interaction, $F\ (4, 96) = 1.65$, $p = .17$. In sum, individuals significantly rate acquaintances with elevated heartrate as more anxious, easily upset, and less calm than those with normal heartrates. In the non-adversarial context, individuals did not rate the acquaintances as significantly more emotional in the elevated condition compared to normal, but this difference was statistically significant in the adversarial context.

The context of interaction (non-adversarial, adversarial) does not have any effect on mood ratings. With clear statistical and practical significance for the overall effect of mood attributions by heartrate condition in both contexts of interaction, Hypothesis 1 is supported.

Our second hypothesis predicts that individuals will make negative assessments about how certain they are regarding the acquaintances' trustworthiness characteristics when the individual has an elevated versus a normal heartrate. We find a statistically and practically significant effect for the heartrate conditions, $F\ (3, 97) = 4.19$, $p < .01$; partial eta squared = .12. However, we also find statistically significant effects for both the context of interaction, $F\ (3, 97) = 2.82$, $p < .05$, and the context x heartrate condition interaction, $F\ (3, 97) = 2.75$, $p < .05$.

A closer inspection of the individual mean differences reveals that the means for all three outcomes (reliability, dependability and trustworthiness) are all lower in the normal condition compared to the elevated condition in the adversarial context (see Figure 2.2). This result is the opposite of what Hypothesis 2 predicts. In the non-adversarial context, we find no statistically significant differences in trust-related evaluations between heartrate conditions. Thus, it is the interaction between the context and the heartrate condition that explains the results: individuals rate acquaintances with normal heartrates significantly lower in terms of trustworthiness, dependability and reliability than those with higher heartrates—but only in the adversarial condition.

Individuals do not rate acquaintances any differently on these three outcomes between the heartrate conditions within the nonadversarial context. In fact, the means for these outcomes are very similar across all conditions and contexts, with the sole exception of the adversarial, normal condition. The mean differences for the trust-related outcomes between the normal and the elevated conditions within the adversarial context are all highly statistically significant ($p < .01$) and highly practically significant: Cohen's $d = 1.1$ (trustworthiness); 1.07 (dependability); 0.68 (reliability). Hypothesis 2 is therefore not supported. However, the strong findings (statistically and practically significant) in the opposite direction from our prediction warrant further exploration in the qualitative results and discussion below.

Qualitative results

Directly after the vignette, participants were asked four freeresponse questions about their reactions to the situation described in the vignette: 1) How do you react to this message, 2) What makes you react this way, 3) What is the ideal outcome of this situation, and 4) What is the worst possible outcome of this situation? The open-field responses were coded into two broad, non-overlapping categories: those that mentioned a negative emotional reaction to the scenario, and those that included a mention of what the other person in the situation might be thinking or feeling. Responses in the latter category were further sub-divided by experimental condition for analysis.

1. Adversarial Context

   This section reports on the qualitative analysis of free responses given by subjects in the adversarial (legal dispute) context.

   a) Normal heartrate

      In the adversarial (legal dispute) context, many subjects who saw a normal heartrate directly indicated that they were negatively adjusting their appraisal of the other person, either in their sympathy toward the other person, or in their judgment of that person's trustworthiness. We find that normal heartrate in the adversarial condition appears to be in conflict with the subjects' expectations about how the acquaintance should feel.

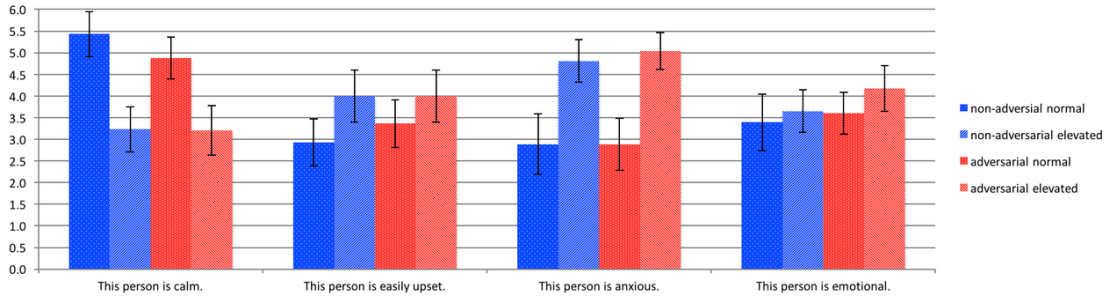

Figure 2.1: Mood-related evaluation means by condition (bars represent standard deviation).

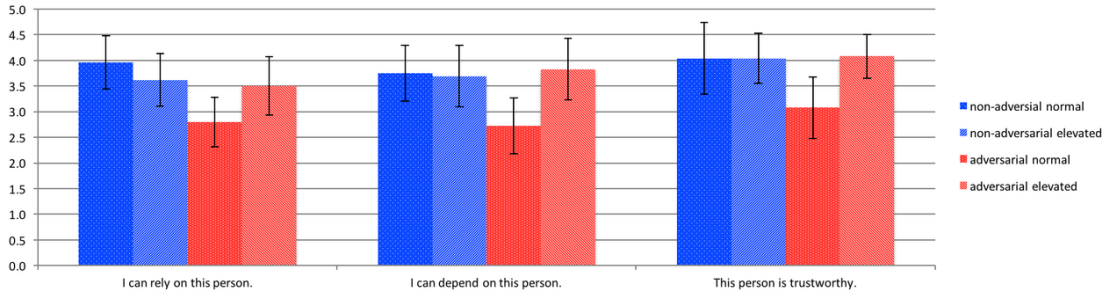

Figure 2.2: Trust-related evaluation means by condition (bars represent standard deviation).

> I will feel less sympathetic to this person because their heart rate doesn't show that they are stressed or upset.
> I feel annoyed because a higher heart rate would indicate that the person cares about the meeting
> The normal heartrate implies that my acquaintance isn't taking this meeting seriously. However, it is difficult to say that my acquaintance does not care or is lying. For example, I have no knowledge of the traffic to determine if my acquaintance is lying.

Here, participants read a lack of care or concern into the acquaintance's normal heartrate, but did not feel the biosignal provided definitive evidence as to whether or not the acquaintance was being truthful. For some participants, however, normal heartrate indicated deception:

> I would think this person is lying. If they were in a rush, their heartrate would be faster.
> I feel like he is lying and is taking his time. I say "hurry up please I can't wait any longer. You are lying to me" It makes me angry to see that his heartrate is normal through all of this. Mine is spiking out of control.

These responses could help to explain the surprising quantitative results of Hypothesis 2 in the adversarial context: the intersection of the adversarial context

with normal heartrate led many participants to view the acquaintance as unsympathetic and, in some cases, disingenuous. As we see below, these negative reactions stand in stark contrast to the interpretations in the elevated heartrate condition.

b) Elevated heartrate

In general, participants in the adversarial context viewed elevated heartrate as a signal that the acquaintance cared about being late.

> Since it shows that the person is trying their best to come, as shown by the elevated heartrate, I would still feel ok.
> I would believe my acquaintance. An elevated heartrate tells me she is probably rushing/hurrying over. I have data from the phone to validate what she is saying to a certain extent.

In these quotes, participants used the elevated heartrate to validate their acquaintance's claim, thus positively assessing their honesty. A few subjects spoke to the power of data in creating what appeared to be objective facts about the other person.

> I won't be angry because seeing this person's heart rate being elevated, it must mean they're in a hurry. Seeing metrics make it easier to believe someone.
> I feel like I'm in a position of power. With the capacity to check someone's heart rate, I can instantly tell how they are feeling. In a way, it is almost like a lie detector.

In both of these quotes, we see attitudes about the presumed authority or "neutrality" of data interacting with beliefs about the body (namely, the relationship between heartrate and emotion, or truthfulness), creating a context in which wearables data can be used to construct social judgments or assessments. How these assessments play out will vary in different social situations, with different sensors, and in different contexts of use. Such variations should be explored much more deeply in future work.

2. Non-Adversarial Context

This section reports on the qualitative analysis of subjects in the non-adversarial context (meeting for a movie),

a) Normal heartrate

In the non-adversarial context, many participants reported that normal heartrate conveyed a lack of appropriate social concern:

> At first I believe that maybe my acquaintance is running late; however, when I discover that their heart rate is normal I wonder why it isn't higher...

> It seems like they are too nonchalant about it
> I feel frustrated because it seems like the person isn't concerned about making me wait.

In these cases, interpretations focused on what the other person was thinking or feeling. As we saw in the adversarial context, normal heartrate seems to be in conflict with expectations. However, unlike in the adversarial context, we did not find evidence that subjects were re-appraising their trust toward the other person. Interestingly, two participants read the normal heartrate positively, as a sign that the other person was telling the truth.

> If his heartrate is normal, then he is probably not lying. I would still be slightly annoyed at this.
> it's OK. her heartbeat was normal, so no lies

These subjects seemed to feel annoyed by the partner's normal heartrate. However, in contrast to the adversarial context, no subjects explicitly stated that the other person seemed less trustworthy, honest or reliable as a result.

b) Elevated heartrate The majority of respondents in the non-adversarial indicated that the elevated heartrate was a token of the other person's regret for being late to the movie. Many participants in this condition indicated that they would have a more sympathetic reaction to the text message as a result of seeing an elevated heartrate.

> Elevated heart rate tells me that the acquaintance at least cares that he/she is late and there's no point in getting mad.
> I would text her back "No problem! I'll grab the tickets and will wait for you out front." It seems obvious she's in a hurry to get there, and is late because of traffic.
> I will feel apologetic because I can see that this person's heartrate is elevated and I do no want him/her to feel worried/ stressed about making a movie.
> I would feel anxiety about being late for the movie and pity because they seem anxious. I don't like being rushed and get anxious when I am rushed

In these responses, heartrate generally seemed to signal that the acquaintance was stressed. While stress is generally assumed to be negative, in this case it seems to engender identification and empathy with the acquaintance. This example gestures toward the highly contextual nature of heartrate's social meaning, and why more work should examine the consequences of these different interpretations.

3. Other interpretations of heartrate: Relevance, validity, creepiness

In addition to the major themes noted above, we also found a few other important interpretations. A small handful of participants (12 total) mentioned aspects other

than the immediate social interaction in relation to the shared heartrate display. The points that surfaced surrounded concerns about privacy, doubts about the accuracy of the sensing device, and doubts about the relevance of heartrate to the particular context.

a) Privacy and disclosure concerns

Only three subjects in the entire experiment pool (n=103) commented on the potential for invasiveness or over-disclosure in heartrate sharing.

> (non-adversarial + normal heartrate) "I feel like I'm violating my acquaintance's private information by knowing their heart beat."
> (adversarial + normal heartrate) "I do suspect the person is lying since his heart rate is normal. I think the extra info of the heart rate is the reason I have a neg. suggestion towards the person. I think the reported heart rate is a bad idea."

Given that heartrate sharing is not (yet) widely deployed in consumer devices, it is somewhat surprising that only a few subjects commented on privacy concerns. This could be partially explained by the fact that the scenario was imagined, rather that simulated, and because subjects might have anticipated our interest in their reactions to the interface.

b) Validity of the device's data

Four subjects mentioned the possibility that the device, or the intuitive inferences drawn from it, may be inaccurate. (adversarial + elevated heartrate) Heart rate could be elevated for many reasons, and just like studies with lie detectors, it may possibly indicate lying, but also could indicate other things. It's just a number, not a definite answer of lying or not. And even then, you've got to forgive people.

> (adversarial + normal heartrate) "The normal heartrate implies that my acquaintance isn't taking this meeting seriously. However, it is difficult to say that my acquaintance does not care or is lying. For example, I have no knowledge of the traffic to determine if my acquaintance is lying. Additionally, my smartphone can be wrong; I don't know how accurate this technology is, especially since it is a very new piece of technology."

Our study did not reference any existing device, so it is possible that the fallibility of particular devices was not on subjects' minds. However, the trust that people place in sensing devices, and the presumed authority of their data, should be explored thoroughly in future work.

c) Relevance of heartrate to the social situation

Only two subjects in the study who mentioned heartrate felt that the data was not necessarily related to the specific social situation described in the vignette:

> (non-adversarial / elevated heartrate) "My initial reaction would probably be to ask them if everything is okay. Their heart rate should probably

not be elevated since they are only driving and weather conditions are not abnormal."
(adversarial / normal heartrate) "There may be reasons why his/her heartrate is normal and why he/she may be late in the first place, so I'm not concerned about that."

Across all conditions, the fact that the vast majority of participants inferred a causal relationship between the heartrate information and the particular social situation highlights the relatively reliable effect of context in priming subjects to draw such inferences. Our results indicate that simply making the heartrate salient, in the absence of other cues, invites people to project a causal narrative on the mood, intentions, and behavior of others.

Discussion

We began this investigation by asking how individuals might interpret heartrate information in uncertain social interactions. Our hypotheses are both based on the simple rationalization that the kinds of negative attributions that people tend to make about their own heartrate will be echoed in their social interpretations of others' heartrates in uncertain contexts. We found, however, a much more complex story about the social interpretation of biosignals and the context of interaction.

Our first hypothesis predicts that an elevated heartrate will be negatively associated with assessments about mood and dispositions in uncertain social interactions, both adversarial and non-adversarial. We found strong support for this hypothesis in both contexts, across our outcome attributions, in line with prior works' findings regarding interpretation of one's own heartrate [**Young1982**]. Our second hypothesis predicts that an elevated heartrate will lead to negative assessments about the partners' trustworthiness, dependability and reliability. As with our first hypothesis, we expected that pre-existing negative connotations with heartrate might translate into negative expectations of trustrelated behavior.

We rejected the second hypothesis in both contexts of interaction. In the non-adversarial context, we found no difference in assessments of trustworthiness, dependability or reliability in the elevated and normal heartrate conditions. Furthermore, we found that the average assessments on these three outcomes were nearly identical between the elevated condition in the adversarial context and the elevated and normal conditions in the non-adversarial context.

Most surprisingly, we find a decrease in trustworthiness, dependability, and reliability in the normal heartrate condition, but only in the adversarial context. As noted in the quantitative results, the differences between the elevated and normal conditions in the adversarial context were highly statistically significant: each of the trust-related measures saw an average decrease of one full point (on a 7-point scale) in the normal condition compared to the elevated condition.

To help explain these results, we turn to our qualitative analyses of the adversarial (legal dispute) context. Subjects in the adversarial context seemed to have expected their partner to have an elevated heartrate. When the partner had a normal heartrate, participants viewed it as evidence that s/he is not bothered enough, not taking the situation seriously, or perhaps even lying. Indeed, many participants explicitly stated in the open text responses that they trusted the partner less because his or her heartrate was normal.

Why do we not see the same effect in the non-adversarial context? Turning again to the qualitative data, we find that participants took elevated heartrate as a token of their acquaintances' genuine desire to arrive on time. It seems that elevated heartrate led many participants in the non-adversarial context to increase their empathy, identification, and understanding of the partners' situation. Thus, even though individuals in the non-adversarial condition associate elevated heartrate with anxiety, lack of calmness, and being easily upset, the negative emotional interpretations do not seem to translate to evaluations of one's trustworthiness, dependability or reliability.

Taken together, we see that heartrate does not inherently (or consistently) affect trust-related outcomes. Instead, social expectations shape interpretations of the heartrate biosignal to create highly contextual, socially-specific meanings. CMC researchers have long noted that, when cues are omitted from technology-mediated interaction, people tend to fill in the gaps [3,10]. However, individuals may interpret new types of interpersonal data in ways we do not yet understand. Our work provides some evidence that such interpretations might have real social consequences. The fact that heartrate alone can significantly alter one's perception of trustworthiness in an adversarial context is an important step towards the larger goal of unpacking social interpretations (and their effects) in technologysupported social interaction. (For one thing, the mostly positive social interpretations of heartrate observed in past work are likely highly dependent on the social context in which they were observed).

Finally, we note a diversity of opinions and interpretations within conditions. For example, a few subjects took normal heartrate as proof of honesty, the opposite view from the majority of subjects. A few subjects did not feel there was necessarily any relationship between heartrate and the social situation at hand. A small minority (three subjects) mentioned concerns around privacy or disclosure. The wide range of views, sometimes contradictory, highlights the complexity intrinsic to interfaces that collect and share biosignals, and warrants future studies into social and contextual interpretation of data from wearable devices.

Limitations

Our vignette experiment examined a single type of scenario in two different contexts, using text-based answers. We still have a limited picture of the range of theoretically important contexts in which individuals may observe and interpret biosignals about others, and a limited understanding of how the rich cues present in realistic interaction contexts might bare on our findings. Our study focused on a first-time interaction with an imagined heartrate sharing interface. We do not know how our findings would hold over time, and it is very likely

that social meanings of any biosignal could become more consistent over time. The vignette scenario was contrived from believable, but currently non-existent smartphone technology. Either due to participants' suspension of their disbelief or due to their actual attitudes about the heartrate sharing, few participants raised questions regarding privacy implications of these scenarios.

Since the vignette study took place online, we could have missed the sorts of rich contextual cues that might be captured by live interviews or other in-person methods. Furthermore, the internet presents a wide array of distractions to survey-takers, and our survey was not able to detect the participants' attention on the task (e.g.., we could not detect whether the subject was switching between tabs in their web browser, or taking breaks during the survey), nor did we monitor how long subjects spent filling out the survey.

While this vignette experiment provides evidence that interpretations of biosignals from sensors (such as wearables) can affect social attributions and behaviors towards others. Nevertheless, many questions remain. A controlled, behavioral experiments could help us ask more specific questions about how elevated heartrate affects perceptions of risk in uncertain interactions, e.g. when money is at stake. Such a study could lead to a more robust understanding of how the transmission of biosignals might affect social behavior. Thus, in the following section, we extend this work to a lab-based experiment.

2.4 Lab-based experiment

Following our vignette experiment, which focused on social attitudes, we extend our inquiry to a trust-building game, which will allow us to study social behavior. Through quantitative and qualitative analyses, we find that "elevated" (versus "normal") heartrate of an exchange partner is associated with negative mood attributions and reduced cooperation in a social dilemma game. To investigate how specific our findings are to heartrate (as opposed to some other "elevated" signal collected from the body), we replicate our initial experiment with an unfamiliar biosignal, "skin reflectivity". We find that both heartrate and the unfamiliar biosignal are associated with negative mood attributions, but we observe a decrease in cooperative behavior only with elevated heartrate. Qualitative results indicate that individuals may learn an association between our unfamiliar biosignal and the cooperative, trusting behavior of their partner. Our findings highlight the role prior beliefs can play in shaping interpretations of a biosignal, while suggesting that designers can, perhaps inadvertently, train users to associate signals with social meanings. We discuss implications for how wearable sensors can mediate social interactions.

TODO remind that we just talked about this in the last exp Generally when individuals believe that their heartrate is elevated, they often believe their mood and emotions to be more negative. Thus, we apply this same logic to how individuals will interpret the elevated heartrates of others in uncertain social interactions:

H1. Participants who see a consistently elevated heartrate from their partner will rate their partner more negatively on mood attributes, compared to participants who see a con-

sistently normal heartrate in uncertain and risky social interactions.

If elevated heartrate has a negative connotation with mood, then elevated heartrate may increase uncertainty about the behavior of one's partner as well. When people know that their partner has an elevated heartrate in an uncertain, risky interactions, they may take actions to protect themselves against potential losses. In trust-building situations, individuals take small risks with other people (entrustment behavior) and learn whether the other person honors that trust or not (cooperative behavior). Thus, individuals have two different ways to respond to increased uncertainty about their partners' behavior in trust situations: 1) reduce the amount they entrust to their partners, or 2) decrease their willingness to cooperate with the partner [**Cheshire2010**, **Cook2005**]. Since we expect elevated heartrate to have pre-existing connotations with negative attributes, we predict that individuals will entrust and/or cooperate less to protect themselves from potential harm when the partner has an elevated vs. a normal heartrate.

H2. Participants who see an elevated heartrate from their partner will (a) trust less, and (b) cooperate less with the partner in uncertain and risky social interactions compared to participants who see a normal heartrate.

Study 1: Sharing heartrate in a risky, uncertain interaction

In order to test our hypotheses, we conducted a repeated trust experiment with shared heartrate information. Trust games present participants with financial incentives to pay attention to their partner's decisions over time, and provide means for operationalizing trust and cooperation in the presence of uncertainty [**Cheshire2010**].

The overall design of the trust game involves anonymous pairs of fixed partners making repeated decisions to entrust valued resources to the partner, and to return (cooperate) or keep (defect) the points entrusted by the other partner. Importantly, individuals can make the highest amount of money when they entrust many points to a partner and the partner returns these points. This creates an uncertain social situation in which participants are trying to earn real money by repeatedly taking risks (entrusting points) to a partner. Since the partners are making the same decisions to entrust and keep/return points from the other partner, these are mutually-dependent social interactions.

1. Experimental Design and Methods

   We operationalized an uncertain social interaction situation using a trust game called the Prisoner's Dilemma with Dependence (PDD) [**Cheshire2010**, **Cook2005**]. The PDD game allows individuals to control the amount of risk that they want to take with their partner by choosing how many points to entrust, followed by a second decision to either keep or return whatever has been entrusted by their partner. Thus, the PDD game separates trust behavior (choosing how much to entrust to a partner) from cooperative behavior (choosing to return or keep what a partner entrusted). In each round of the PDD game, participants were given an initial endowment of 10 points. Each participant decided whether to entrust any number of points to their partner,

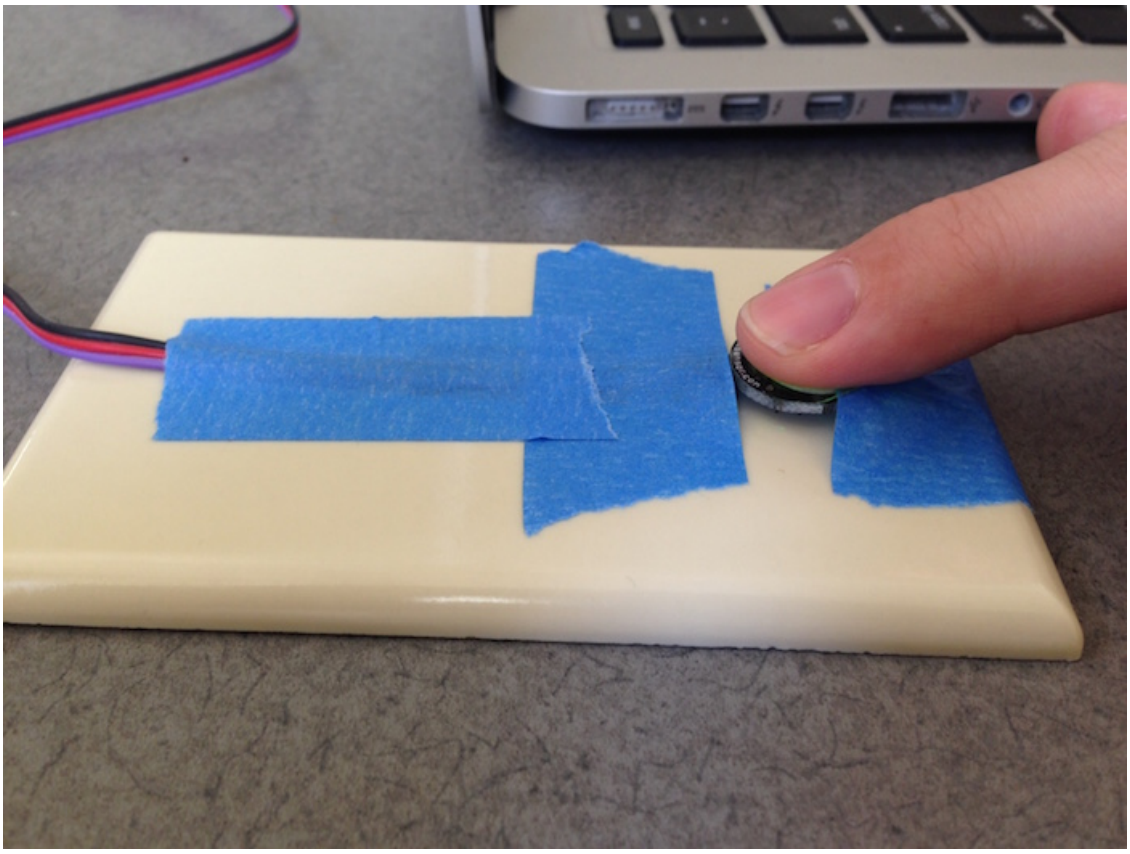

Figure 2.3: The heartrate monitor. Participants were told to place their finger on the monitor to take a reading while viewing their partner's decisions during the previous turn.

from zero to ten. Then, participants found out at the same time whether their partner had entrusted them with any of their own points, and if so, how many. Next, each participant decided whether to keep the points entrusted to them (defection) or return them (cooperation). The participants could not return only a portion of the entrusted points, only all or none of them. If the points were returned to the partner, they were automatically doubled in value for that participant.

After all participants made decisions about returning or keeping any points that had been entrusted to them, they were then asked to place their finger on the heartrate monitor for a few seconds in order to get a pulse reading (Figure 2.3). Participants then viewed the summary of point calculations for the round. Subsequently, participants viewed a visual display of the partners' recent heartrate (Figure 2.4). The final point calculation for the round included any of the initial allotment of points remaining after the trust decision, plus and points that the participant kept from their partner if they decided not to return them. In addition, players received points for any entrusted points that their partner returned, which doubled in value.

When participants arrived at the laboratory, they were given a consent form that described the nature of the study, as well as the human subjects' approval information from our university. We wanted participants to believe that they would be interacting with other real people, and this perception was enhanced by having 12-16 participants at separate computer terminals in the same large room during each experimental session. In fact, we controlled the trust and cooperation behavior of the "partner" for every participant using a simulated computer actor. As a result, no one in the study interacted with a human partner.

The simulated actor was programmed to always begin by entrusting one point on the first round, then randomly entrust up to one point above or below whatever the partner entrusted on the previous round. In addition, the simulated actor was programmed to always cooperate (i.e., return the points that were entrusted by the partner). Following **[Cheshire2010]**, we chose to use a highly cooperative interaction partner in order to minimize any other forms of uncertainty in the interaction. A highly-cooperation partner does not introduce any defection behaviors that might otherwise reduce cooperation or trust from the participant (thereby hindering our ability to detect main effects from the experimental manipulation). Thus, the simulated actor was designed to reciprocate the entrusting behavior of the human participant on each round, and always cooperate no matter what the human participant chose to do.

The participants completed 20 rounds of the PDD game, but they did not know how many rounds they would play in order to eliminate end-game effects. After all rounds of the PDD game were completed, participants answered a short post-questionnaire in order to assess their attitudes and beliefs about their partner. This questionnaire included 7- point Likert-style response questions (1 = strongly disagree, 7 = strongly agree) about the partners' beliefs about the partners' anxiety (e.g., "my partner is anxious" and "my partner is calm").

As a manipulation check on the perceptions of the simulated actor's behavior, we also asked questions about the partners' game behavior ("my partner is trustworthy" and "my partner is cooperative"). Finally, we supplemented our quantitative measures with two open-ended questions: "How would you describe your partner?" and "What, if anything, did heartrate tell you about your partner during this experiment?" Participants were paid between $15-30 based on their point earnings during the game. The entire study lasted one hour.

At the end of the study, participants were debriefed on the true nature and intent of the experiment. An experimenter was available at the end of the study in case of any questions, and we provided participants with the researchers' email addresses on both the signed informed consent form, as well as the debrief form, so that they could contact us regarding any aspect of the study. We did not receive any emails or concerns from participants.

2. Experimental Manipulation

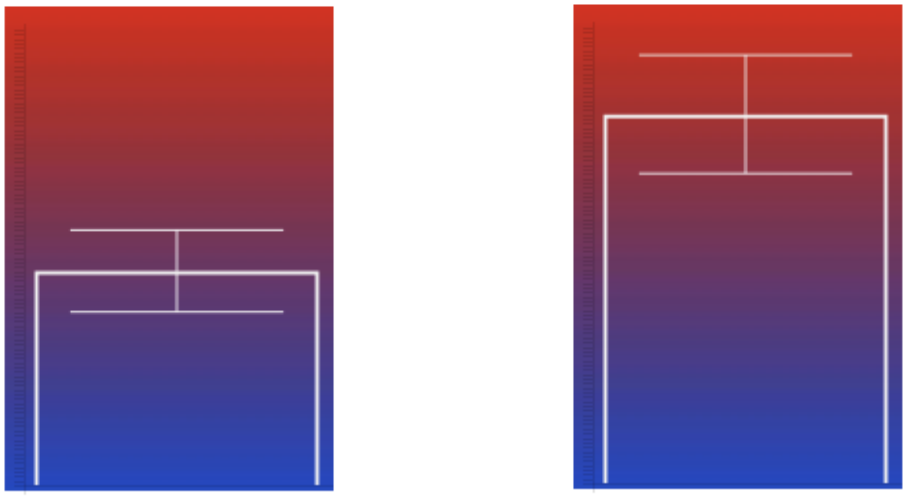

Figure 2.4: The heartrate visualization. After viewing the results of the previous round, participants saw a graph of what they believed to be their partner's heartrate, either normal (left) or elevated (right). Error bars fluctuated within pre-set bounds.

To assess the effect of interacting with a partner who has an elevated heartrate versus interacting with a partner who has a normal heartrate, we controlled the heartrate information that participants saw after each round of the experiment. This created a two-condition design: always normal heartrate (NH) and always elevated heartrate (EH).

3. Participants and Procedure

   Our sample was undergraduate students recruited from the population of a large west coast public university in the United States. We contacted potential participants via email from a voluntary experimental subject pool. All participants expected to be contacted to participate in a social research study at some point during the semester, and knew that they would earn between $15-30 during this one-hour study, depending on their choices during the experiment. Fifty-six participants (56) completed the experiment, 41 women, 14 men, and one self-identified as other. The mean age of participants was 21.

   Upon arrival at the laboratory, participants were guided to an individual desk with privacy walls. After signing an informed consent form, participants read written instructions on the computer which explained that they will have the opportunity to interact with a single partner for many rounds in order to examine decision making in social situations. Participants were also told that we would collect pulse (heart rate)

information at designated times during the study using a simple pulse monitor that was connected to the laptop computer.

4. Validity Check of the Visualization

   Our study aims to understand the effect of "elevated," as compared to "normal," heartrate. As such, we needed to show participants a visualization that afforded only a relative value for heartrate, not an exact figure (since different people may have different ideas of what number value constitutes a normal or elevated heartrate).

   We designed a visualization to display a relative heartrate (Figure 2.4) and performed a small (n=25) face validity check to ensure that our visualization would work as intended in the actual experiment. In our short validity survey, we included three versions of the visualization, representing a mix of elevated, low and normal heartrate, and two Likertscale questions: "The precise meaning of this graphic is ambiguous," and "I can interpret the difference between 'low', 'normal', and 'high' heartrate from this graphic," which participants answered from "Strongly Agree" to "Strongly Disagree" on a 5-point scale. We also included two open-ended questions, "Please explain what the picture is telling you about one's heartrate," and "Please explain what this picture does not tell you about one's heartrate."

   We distributed this survey over an email list to students and alumni of a public, West Coast US university, and received 25 valid responses. The answers to both Likert questions indicated agreement that the visualization was both ambiguous (mean = 3.58, S.D. = 1.28) and also easily interpretable (mean = 3.41, S.D. = 1.35). Importantly, openended qualitative responses confirmed that the heartrate was easily understandable, but that the precise value of heartrate was ambiguous.

Study 1: Results

1. Quantitative results

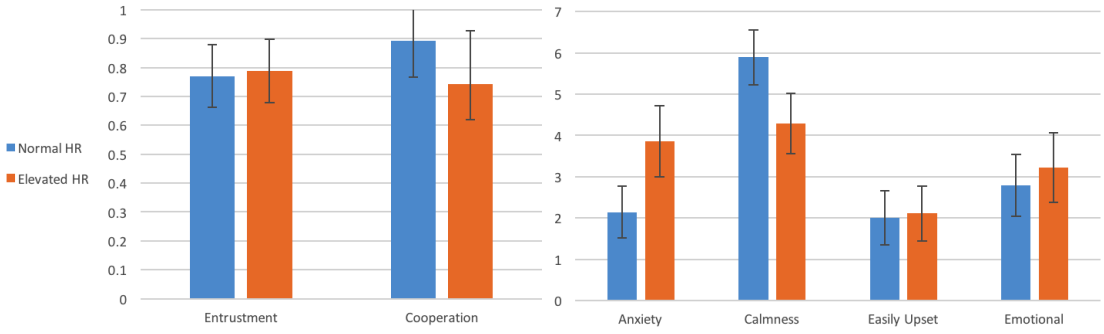

Figure 2.5: Means of entrustment and cooperation (left) and mood attributions (right) in elevated and normal heartrate conditions.

Our first hypothesis predicts that, when individuals believe that their partner has a consistently elevated heartrate, compared to a normal heartrate, they will rate the partner more negatively on mood attributes. Consistent with prior research, we found an overall strong, statistically significant effect and medium practical association between attributions and experimental condition, $F(4, 51) = 6.7$, $p < .0001$; Wilk's lambda = .66, partial eta squared =.34. Turning to the individual outcomes, we find that perceptions of the partners' anxiety is significantly higher in the EH condition ($M = 3.86$, $SD = 1.72$) compared to the NH condition ($M = 2.14$, $SD = 1.27$), $F(1, 54) = 18$, $p < .001$; partial eta squared = .25. Furthermore, participants rated their partners as significantly more calm in the NH condition ($M = 5.9$, $SD = 1.3$) compared to the EH condition ($M = 4.29$, $SD = 1.46$), $F(1, 54) = 18.71$ $p < .001$; partial eta squared =.26. On the other hand, we found no statistically significant differences for perception that the partner is "easily upset" or that the partner is "emotional" (p = n.s.). In sum, we find strong statistical and practical differences in perceptions of both anxiety and calmness, but no statistical or practical differences in perceptions of how emotional or easily upset the partner is in the two experimental conditions. Given the significant omnibus test and significant results on two of the four individual outcomes, Hypothesis 1 is partially supported.

Our second set of hypotheses predict that participants in the elevated heartrate (EH) condition will exhibit lower trusting (H2a) and/or cooperative (H2b) behavior compared to those in the normal heartrate (NH) condition. The average points entrusted by participants in the EH condition ($M = 7.88$, $SD = 2.18$) was not significantly different than the NH condition ($M = 7.7$, $SD = 2.18$), t =.28, p=n.s, one-tailed test. Thus, individuals entrusted points to their partners at approximately the same level in both conditions (Figure 2.5). Hypothesis 2a is not supported.

However, we found that the average cooperation rate in the EH condition ($M = .74$, $SD = .37$) was statistically significantly lower than the NH condition ($M = .89$, $SD = .25$), $t = 1.76$, $p < .05$, one-tailed test. Importantly, this result shows a medium practical effect size (Cohen's $d = .47$), indicating a meaningful real world difference. On average, those in the normal heartrate condition cooperated 20% more than those in the elevated heartrate condition (Figure 2.5). Hypothesis 2b is supported.

a) Manipulation Checks

Since we designed the simulated actors in both conditions with trusting and always-cooperative behavior, we did not expect participants to rate the simulated actors differently in terms of the focal behaviors of cooperativeness and trustworthiness between experimental conditions. This is a critical manipulation check, since we need to rule out any perceived effect of the simulated partners' behavior in order to establish that the primary treatment (heartrate of partner) had an effect on the human participants' behavior. The omnibus test of difference in perceptions of the trustworthiness and cooperative behavior between conditions

was not significant, F(2, 53) = .21, p = n.s.; Wilk's lambda = .99, partial eta squared =.01. Thus, as we would expect, individuals did not indicate significant behavioral differences for the trusting, cooperative simulated actor (which was programmed to behave exactly the same in both conditions).

2. Qualitative results

At the end of our questionnaire, before the demographic questions and the debriefing, participants were presented with two open-ended questions. The first asked participants to "Tell us how you would describe your partner." The second asked participants "What, if anything, did heartrate tell you about your partner during this experiment?" This section discusses and unpacks some of the responses that these questions elicited.

a) Elevated Heartrate

Many people who referred to elevated heartrate in their responses mentioned that it signaled anxiety. In some cases, participants even reflected on a negative relationship between elevated heartrate, anxiety and trust:

> how excited he/she is, whether he/she cheated
> It was elevated all the time so I think s/he was anxious [...] so I guess s/he did not completely trust me

These quotes further support our first hypothesis, as well as findings of past work showing that elevated heartrate typically signals anxiety and mood. In other words, elevated heartrate (and heartrate in general) seemed to be about the partner's current disposition, rather than who the partner was as a person. While the majority of those who mentioned elevated heartrate implied a causal relationship between the signal and the game context, a few did not:

> My partner's heart rate was elevated the whole time, most students are stressed so that might be why.
> They may have been nervous because of doing the experiment itself.

The relative rarity of skepticism about the relationship between heartrate and specific game events highlights the crucial role of framing and salience in turning what might be a disembodied signal (heartrate data) into a relevant, contextual clue. We also noted diversity in beliefs about the meaning of heartrate itself. Where almost all participants who mentioned heartrate associated it with anxiety, at least one participant had an entirely different take on his/her partner's consistently elevated heartrate:

> My partner's heart rate does not change too much which indicates that he or she is very nice.

These quotes highlight overall diversity in what an elevated heartrate is capable of meaning. Even within our relatively small, and relatively homogenous sample

of university students, our quotes imply a mostly negative association with elevated heartrate, but also a potentially long tail of diverse beliefs about elevated heartrate.

b) Normal Heartrate

Many participants said that normal heartrate indicated that the partner was "calm," "chilled out," or "not anxious." [HR signaled] that my partner was always calm. The heart rate never fluctuated, it didn't make a difference.

> They remained calm
> I think it showed that my partner wasn't too nervous to see if he/she was returned the points or not, maybe because it was just an experiment or maybe because he/she wasn't worried about what result he/she was about to see was.

These quotes show subjects inferring a direct connection between the heartrate signal and the attribution of a calm mood. One participant specifically mentioned that consistency of normal heartrate made their partner seem more trustworthy:

> My partner's heart rate has been consistently normal throughout the experiment, so I guess s/he has no intention to cheat.

Another participant, presumably a cooperative one, thought that their partner's heartrate would have risen if s/he had not cooperated:

> I think it remained the same [normal] because I paralleled my partner's actions whereas if I had contradicted them, their heartrate probably would have changed in response.

In all of the above quotes (and the vast majority of responses), participants inferred a relationship between normal heartrate and calmness. However, a few participants did not infer any relationships between behavior, moods and the signal they saw.

> Heartrate did not tell me anything. My partner was average each time.
> I also am sure I have an elevated heart rate due to coffee consumption so I did not take my partners into consideration.
> I based my decisions on their previous actions.

Not every participant explicitly inferred a calm mood from the normal heartrate signal, but most did. Taken alongside our quantitative results, our qualitative results provide evidence that subjects have used the emotional attributions they made based on their partner's normal heartrate to guide their behavior in the trust game.

Study 2: Sharing an unknown signal in a risky, uncertain interaction

In study 1, we found that participants cooperate less with partners who have elevated heartrates in the repeated trust game, compared to those with normal heartrates. While this result supports one of our key hypotheses, it also begs another question: Is the effect we observe due to heartrate specifically, or might any elevated biosignal show the same results for negative perceptions of mood and reduced cooperative behavior towards the partner?

In our second experiment, we attempt to tease out the effect of the heartrate signal itself, compared to any "elevated" (versus "normal") signal collected from the body. We replicate the first study, except that we tell participants that our monitor device measures SRI (Skin Reflectivity Index). SRI is an unfamiliar biosignal, for which individuals should not have any prior cultural or social beliefs.

1. Hypotheses

   Without any context for what SRI means as a signal, participants may assume that any biological signal that is "elevated" from normal will be negatively associated with one's mood. If this is the case, then we should observe the same general pattern of negative mood attributions and less cooperative behavior when the partner has an elevated SRI as we observed with heartrate.

   On the other hand, perhaps heartrate is special due to its common social associations with mood, anxiety, and even deception. If heartrate is distinctive in this regard, then we would not observe the same significant differences between normal and elevated SRI and mood attributes, trust, and cooperation rates with the partner.

   To test the effect of our unfamiliar biosignal on behavior in risky, uncertain interactions, we evaluate the exact same hypotheses from study 1 again in the context of SRI: H3. Participants who see a consistently elevated SRI from their partner will rate their partner more negatively on mood attributes, compared to participants who see a consistently normal SRI in uncertain and risky social interactions.

   H4. Participants who see an elevated SRI will have lower (a) trust rates (b) cooperation rates in uncertain and risky social interactions compared to participants who see a normal SRI.

2. Experimental Design and Methods

   The second study was identical to the heartrate study in every way, except that we told participants we were measuring "Skin Reflectivity Index," instead of heartrate. All mentions of the word "heartrate" in our original experiment software were replaced with "SRI" and/or "Skin Reflectivity Index". We purposely did not define or explain what the SRI signal is, or what its measurements mean. All participants were debriefed on the true nature of the experiment at the conclusion of the study. This debriefing included the fact that the partner was based on idealized behavior, and "SRI" was

actually just a term for heartrate, as collected by a standard light-based pulse sensor. As with the first study, participants had the ability to ask the experimenter questions at the end of the study, or send an email if they had additional questions or concerns. We did not receive any follow-up concerns from participants. The only other variation from the first experiment is that, in the SRI experiment, we told participants to place their palms an inch above the light sensor rather than to place their fingers on the monitor. Since placing a finger on a light sensor is a familiar of measuring heartrate, this was done to reduce the possibility that participants would think that SRI is actually heartrate.

3. Participants

   We recruited our sample for the second study from the same population and using the same method as described in study

   a) Our recruitment procedures ensured that no one who

   participated in the first study could be recruited for the second study. Sixty-three participants (63) completed the second experiment, 40 women, 22 men, and one self-identified as 'other'. The mean age of participants was 21. Importantly, the gender distribution and age of the sample was equivalent to the first study.

Study 2: Results

1. Quantitative results

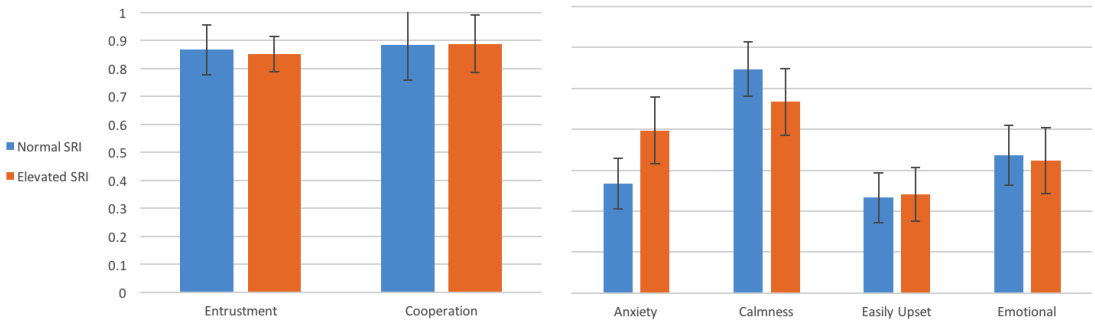

Figure 2.6: Means of entrustment and cooperation (left) and mood attributions (right) in elevated and normal SRI conditions.

H3 predicts that when individuals believe that their partner has a consistently elevated SRI, compared to a normal SRI, they will rate the partner more negatively on mood attributes. As with the first study on heartrate, we found an overall strong, statistically significant effect and medium practical association between attributions and experimental condition, $F(4, 59) = 4$, $p < .01$; Wilk's lambda $= .79$, partial eta squared

=.21. For the individual outcomes, we find that perceptions of the partners' anxiety is significantly higher in the elevated SRI condition (M = 3.97, SD = 1.62) compared to the normal SRI condition (M = 2.67, SD = 1.24), F(1, 62) = 12.8, p < .001; partial eta squared = .17. Furthermore, participants rated their partners as significantly more calm in the normal SRI condition (M = 5.5, SD = 1.3) compared to the elevated SRI condition (M = 4.68, SD = 1.63), F(1, 62) = 4.4 p < .05; partial eta squared =.07. Just as with the heartrate study, we found no statistically significant differences for perception that the partner is 'easily upset' or that the partner is 'emotional' (p = n.s.). In sum, we find strong statistical and practical differences in perceptions of both anxiety and calmness, but no statistical or practical differences in how emotional or easily upset one perceives the partner to be in SRI conditions. Given the significant omnibus test and significant results on two of the 4 individual outcomes, Hypothesis 3 is partially supported.

Our final hypotheses predict that participants in the elevated SRI condition will exhibit lower trusting (H4a) and cooperative (H4b) behavior compared to those in the normal SRI condition. The average points entrusted by participants in the elevated SRI condition (M = 8.5, SD = 1.27) was not significantly different than the normal SRI condition (M = 8.7, SD = 1.77), t =.39, p = n.s, one-tailed test. Thus, individuals entrusted points to their partners at approximately the same level in both conditions (Figure 2.6). Unlike the heartrate study, however, we found no significant difference in cooperation rate between in the elevated SRI (M = .89, SD = .21) and the normal SRI condition (M = .88, SD = .25), t = .09, p = n.s., one-tailed test. H4a and H4b are not supported.

a) Manipulation Checks

As with the first study, the simulated actors in study 2 were programmed to be consistently trusting and cooperative in the elevated and normal SRI conditions. Thus, we do not expect participants to rate the simulated actors differently in terms cooperativeness and trustworthiness between experimental conditions. As expected, the omnibus test of difference in perceptions of the trustworthiness and cooperative behavior between conditions was not significant, F(2, 61) = 3, p = n.s.; Wilk's lambda = .91, partial eta squared =.09.

2. Qualitative results

As in the heartrate condition, participants in the SRI condition were asked open-ended questions at the end of the post-experiment questionnaire, before the demographic questions and debrief. As in the heartrate condition, participants were asked how they would describe their partner. However, unlike in the heartrate condition, participants were asked, "Recall what we were measuring with the sensor. Please describe it below." After completing this question, participants proceeded were given two more open-ended items: "What, if anything, did SRI (skin reflectivity) tell you about your partner during

this experiment?" and, "As a signal, what do you believe that SRI says about another person?"

a) The Meaning of an Unfamiliar Biosignal

We purposely did not explain what SRI might mean in this study. Nevertheless, when asked what was being measured in SRI, some participants gave us thorough explanations: The "reflectivity" part of SRI leads me to believe that the device is measuring how much light is reflected by a person's palms, which leads me to assume that SRI is increased when a person's hands are sweatier, and thus more covered in water, which reflects light better than simply someone's skin.

While explanations like this one indicate that participants believed our signal was real, reports of what participants thought SRI meant in the context of the game are more relevant to our analysis here. Like in the elevated heartrate conditions, and elevated SRIs were associated with either nervousness or excitement.

> If the SRI reads high, it may indicate that the person expects to be betrayed in some way or is hopeful of a positive result. I forgot what SRI stands for again. Since his/her SRI is always elevated, I would assume he/she is nervous/excited or just it's hot in here.
>
> SRI may give insight as to how nervous or excited someone's response is to something that happens. Maybe someone with a larger range in SRI is more emotional.

These assessments of SRI are quite similar to interpretations from the elevated heartrate, and corroborate our quantitative findings that those who saw elevated SRI rate their partners as more nervous. However, the fact that these emotional assessments were similar in both elevated heartrate and elevated SRI conditions, but behavioral outcomes were different, challenges our notion that negative emotional cues caused these behavioral outcomes—a point we address in more detail in the discussion below. As in the heartrate conditions, some participants responded that SRI told them little or nothing of interest about their partner:

> Nothing at all about the person other than an arbitrary value of a sensor. Since the SRI seemed to be bouncing around in the blue range but never got into the red range (which I assume would be "abnormal" since the blue range was normal) I don't think SRI is an accurate measurement of much. As with heartrate, people cannot always be convinced that a biosignal is informative, even after many rounds of conditioning and a highly suggestive context. However, as in the heartrate condition, responses indicating that SRI had no meaning were a clear minority in our sample.

b) Elevated SRI

To help explain why elevated heartrate had a chilling effect on cooperative behavior, where elevated SRI did not, we delve into the responses of participants in

the elevated SRI condition. When asked what SRI told them about their partner, participants often reported nervousness or anxiety, just as we noted in the quantitative results:

> [SRI shows] stress or heightened anxiety
> how reactive they are, or how close to the surface their emotions are.
> The nervousness of a person.

However, we noted that a significant number of participants in this condition mentioned that elevated SRI had some sort of positive association with behavior—even though it is also interpreted as indicating anxiety.

> Elevated means they feel safe and trustful. Lower than average means they are defensive and scared.

This interpretation stands in stark contrast to elevated heartrate, which also signaled anxiety, but had a negative association with behavior. In explaining why participants found elevated SRI to signal cooperativeness and trust, we look toward the responses of participants who seemed to learn a meaning for this signal:

> Well, since their SRI was always high and they always gave the money back to me, (based on these only two bits of info I know) I assume the two are correlated and an elevated SRI means that they're going to give the money back. [...] I guess it means that they're trustworthy and will do the right thing by their partner.
> I cannot tell [what SRI means], but my partner's was extremely elevated for the whole experiment and s/he was good at conducting mutually beneficial transactions.

These quotes strongly suggest that, unlike for heartrate, SRI participants picked up on a pattern between their partner's always-cooperative behavior and the elevated biosignal that we displayed to them, thus filling in the gaps about what SRI meant in this context. In contrast, we found no evidence that elevated heartrate participants learned such an association in the first study, despite the fact that every participant interacted with a perfectly cooperative partner in all conditions and studies.

c) Normal SRI

As with those in the elevated SRI condition, many participants in the normal SRI condition identified some relationship between SRI and the other person's mood. I think this helps identify how people are feeling internally when making decisions.

> his/her mood at that point of time
> [SRI shows] stress or heightened anxiety
> how anxious they are.
> I think our anxiety is being measured.
> How anxious/nervous someone is, if their SRI is high

> In some cases, participants in the normal SRI condition inferred that elevated SRI might have a negative meaning: not to sure, high sri may indicate panic/fear or anger low sri may indicate calmness and contentness. A person is less likely to trust other people if he or she has a high SRI.
>
> Overall, the responses for both SRI conditions support the interpretation that participants learned an association between cooperative, trustworthy behavior from the partner and SRI. As we argue in the following discussion, such associations are more likely in the SRI conditions because, unlike for heartrate, participants should have no preexisting beliefs or associations with SRI.

Limitations

Controlled, laboratory studies always come with clear advantages (such as high internal validity) and disadvantages (such as reduced external and ecological validity). Our study did not attempt to emulate a real-world interaction context with a biometric sharing device, though this is a clear next step, now that we know there are important differences in how biosignals are interpreted. Furthermore, our use of highly cooperative, computercontrolled interaction partners with stable biosignals (always high or always normal), prevents us from being able to speak to the effects of more dynamic behaviors and/or changes in biosignals over longer periods of time. From these experiments, we also do not know how these results will transfer to other contexts, and other types of social interactions. Also, our study by nature focused on first-time, iterated interactions, both with an interface and with another unknown person. We do not know how these results might apply over the course of more personal relationships, or after repeated experiences with a specific interface in a biosignal sharing device. In addition, this research was conducted on young adults at a large public university, which is an important limitation when considering whether these results would hold across age groups and other key sources of sociodemographic variation in the larger population.

2.5 Discussion

We found that both heartrate and SRI signaled negative mood to participants, including anxiety and lack of calmness. It is possible that almost any "elevated" biosignals could be associated with negative mood attributions such as anxiety and lack of calmness: many elevated signals (pulse, temperature, blood pressure) carry associations with being angry, sick, hot-headed, and a host of other negative attributions. People may default to such attributions when seeing an unknown signal that comes from the body.

Elevated heartrate had a chilling effect on cooperation, where an unfamiliar biosignal, SRI, did not. So, why did the negative mood attributions in the elevated SRI condition not translate into reduced cooperation, as they did for elevated heartrate?

Our results shed light on two relevant phenomena that may address this question. First, pre-existing beliefs about heartrate are powerful: even when playing with a very cooperative, trusting game partner, negative connotations surrounding elevated heartrate appear to lead individuals to cooperate less. Our results suggest that participants bring to uncertain social interactions their own expectations about what elevated heartrate means, and that these biases cannot be quickly overridden, even when behavioral evidence sends a positive message (e.g., high cooperation and trust from the partner).

Second, we find evidence that participants can "learn" a social meaning for a previously unknown signal. Our qualitative data suggest that participants in the SRI condition associated whichever signal they saw (elevated or normal) with cooperativeness, and trustworthiness. Unlike with heartrate, people did not have preconceived notions of how SRI should affect the social behavior of the partner, since SRI does not exist. Instead, we observe participants discovering "what SRI means" by watching their partner's behavior in relation to the biosignal. In the absence of guidelines for interpreting what SRI is or what it measures, individuals appear to fill in the gaps with available behavioral information.

If people can learn social meanings for previously unknown signals, perhaps even pre-existing connotations for familiar biosignals could change over time. After all, the meanings of a signal like heartrate are the product of associations that have been shared and developed over centuries. However, technology allows for new expressions of these ancient signals [**Slovak2012**]. If social heartrate information became an easily accessible biosignal in trust-based interactions like negotiations, we might find its social meaning could evolve further. Unfortunately, short-term laboratory studies such as this one are unlikely to trigger or detect enduring shifts in the social meanings of familiar biosignals. We need both longer-term experiments, and mixed-methods research that can draw from rich qualitative data as well as statistically and practically significant changes in interpretations over time.

Broadly, our results raise questions about how and why unfamiliar signals take on social meanings in different contexts of interaction. Researchers in CSCW and HCI have long noted our tendency to read into cues and signals in technology-mediated communications. From impact factors and citation counts in scholarly work [**Elsden2016a**] to societal indices [**Wilson2003**] , to health metrics such as the bodymass index (BMI) [**Campos2004**], human have a tendency to impart "real" meanings onto metrics, scales and signals – meanings that may not align with the concepts their designers aimed to measure. It is critical that we continue to question how biosignal data could shape our interpersonal interactions, and whether the outcomes will always translate into meaningful social information.

2.6 Implications for design

From research projects like the sociometer, which produce "social metrics" [**Wu2008**], to consumer devices like the Spire, which compute "calmness" or "focus" quotients [**SpireInc**], developers are throwing different biometric signals at people faster than they can learn what the signals mean in context. In the absence of strong cultural beliefs about the signal, people

could produce correlative assumptions similar to the ones we observed in our experiment. Designers should take care to establish what the signals in the applications mean, or could mean. Testing the limits of what people are willing, or able, to believe, and whether these beliefs transfer between different contexts, could have wide-reaching implications for those who design interactions with wearable biosensors.

On the other hand, many research and commercial projects use signals that people might associate with commonly understood experiences (e.g. a racing heart, a sweaty palm). Designers should strongly consider how these embodied experiences might color the conclusions that users jump to, and bound what users are willing to believe.

We also hope that researchers will investigate settings in which biosignals vary over longer time periods, perhaps with a more naturalistic technology probe study. Such a study could help us understand how prior beliefs about signals both affect and are affected by social interactions in the course of everyday life.

In general, wearable sensors can enable social interactions in which we share more information than is normally possible face-to-face. The ability to surface signals that are normally socially invisible (e.g. heartrate, or galvanic skin response) presents new territory for designers of computermediated interactions. While recent work has explored how these novel signals fit into our existing understanding of social cues [**Howell2016**], much work remains.

2.7 Conclusion

We find that sharing heartrate can negatively influence trusting attitudes and behaviors. However, heartrate alone does not communicate trust. Instead, individual's social expectations interact with the heartrate data to produce context-specific meanings. Complicating matters further, our qualitative data reveal a diversity of interpretations regarding the relevance and meaning of a heartrate in context, and the privacy implications of biosensing technologies. Our findings advance and complicate our understanding of the role that biosignal sharing can play in social, computer-mediated contexts, and motivate more detailed study into the mechanisms by which social interpretations arise from basic physiological signals.

Further, our experimental results imply that interfaces can "teach" the meaning of some biosignals, where others carry strong, pre-existing connotations that even repeated interactions cannot easily alter. In general, prior beliefs about the body (drawn from culture, lived experience) seem to shape what a biosignal can mean in a given context. However, in the absence of prior beliefs, there exists an opportunity—and a potential danger—that designers of biosignal-sharing systems can condition participants to learn (potentially arbitrary) associations between biosignals and social behaviors.

Aside from heartrate, we do not know what many other biosignals might be associated with moods and behaviors. Other biosignals (e.g., galvanic skin response, electroencephalography or EEG), could offer different affordances for sense-making. It is unclear from our work how the social interpretation of the signals from these devices could affect social behaviors

such as dyadic and group trust. Similar studies with signals from, e.g., the brain [**Ali2014a**] are a clear direction for future work. Especially interesting cases are signals for which precise or empirical meanings are still being hotly debated, such as EEG (brainwaves), a sensing modality we begin to discuss in the following chapter.

Chapter 3

TODO Shifting to the brain

While the prior chapter establishes that people build mind-related meanings around biosensory data, this chapter locates brainscanning, and brain-based authentication specifically, as a fruitful case for understanding how particular sensing technologies construct notions of mind. I report on the qualitative and quantitative results of survey among participants in a large (n>10,000), longitudinal health study, and an Amazon Mechanical Turk population.

3.1 TODO Introduction

TODO Last study looked deeply into heartrate

TODO Now time to look at all the diff sensing modalities

3.2 TODO Methods

The survey we report on here, currently in-progress, examines how people's beliefs differ given device ownership, and their membership in one of two groups: Mechanical Turk workers, or people enrolled in Health-e-Heart, a massive (n > 40,000), longitudinal study, in which volunteers fill out surveys about themselves, and/or upload data from biomedical self-tracking devices, over the course of several years [**Estrin2010a**].

TODO Motivation behind subject choice

1. **TODO** why heh?
2. **TODO** why mturk?

Subjects / Survey

In one portion of the survey, we ask subjects to rate a number of different biosensors in order of how likely individual's believe each sensor is to reveal what "a person is thinking or feeling" (Figure 3.1). This section reports on a subset of Mechanical Turk workers (n=100) and Health-e-Hearth subjects (n=100).

3.3 TODO Quantiative results

In our preliminary findings, brainwaves (EEG) are seen as among the most revealing biosignals, just below body language and facial expression, in their capacity to reveal the inner workings of a person's mind. More common sensors such as GPS and step count are seen as less revealing (despite empirical evidence suggesting such data can be quite revealing indeed [**Canzian2015**]).

3.4 TODO Qualitative results

Our qualitative data revealed that subjects in both groups generally believed EEG to reveal various details about the mind, mood, emotions, and identity. We asked subjects to reflect on why they answered the way they did during the ranking task (Figure **??**). In the Health-e-Heart group, several subjects gave relatively specific explanations as to why they ranked EEG hihgly.

> *(S24) I assume some information can be gleaned from brain wave activity in various parts of the brain related to rewards or executive control, but without accompanying information, it may be difficult to discover my thoughts.*
>
> *(S23) EEGs note parts of the brain that are active. Again, in conjunction with other measurements, I suspect that some sense of what one is thinking and feeling could be learned.*
>
> *(S91) I would rate this relatively high on the list because science has shown that we can detect a lot about which areas of the brain are accessed and at which times. This can tell a person a lot about what they might be thinking and especially how they are feeling.*

While these explanations range somewhat in their specificity and confidence, they share the general sentiment that EEGs can be revealing. Subjects in the Mechanical Turk condition broadly shared this belief, though tended to use less physiological detail in their explanations.

> *(S157) Brain activity can pinpoint exact emotions by monitoring certain areas on the brain.*

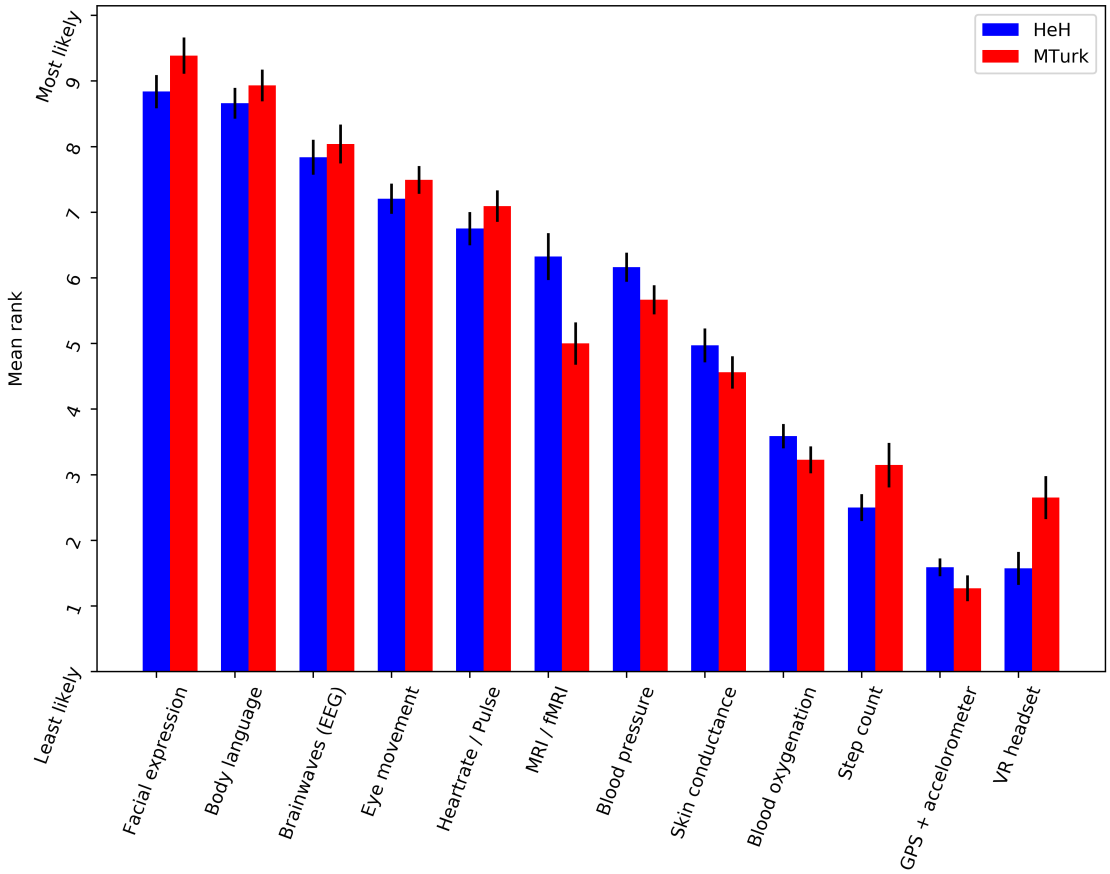

Figure 3.1: "Please rank the following sensors in how likely you believe they are to reveal what a person is thinking and feeling." Higher bars indicate higher rank, or higher likelihood of being revealing.

> *(S130) Brainwaves could tell you a lot more about what someone is thinking and feeling. You could measure the patterns of brainwaves in an experiment.*

Meanwhile, some subjects from both groups did not fit this trend. Ten subjects ranked EEG low in its ability to measure what a person is thinking or feeling. Their qualitative answers revealed a diverse set of reasons for this ranking. Three subjects indicated a general lack of faith in brainwave's reliabilty.

> *(S20) I don't think we have the ability to translate brainwaves into thoughts or emotions.*
>
> *(S101) EEG is very nonspecific and rarely can tell details reliably.*
>
> *(S138) Possible but not accurate.*

These explanations broadly centered around EEG as a signal. They range somewhat in their confidence, from a fundamental skepticism (S20) to caveats about possible accuracy or specificity (S101, S138). In contrast to these three subjects, S10 ranked EEG low because s/he felt the premise of a consumer grade EEG was implausible.

> *(S10) I assume that scientists can identify by brain patterns what others are feeling and thinking based off of years of research. I've never heard of a consumer grade eeg - and doubt it could be as powerful as a laboratory eeg. If it is then I would be interested in this product.*

This subject's explanation surfaces the practical differences in attitudes that people might have to a technology's theoretical existence, and its realized existence as a consumer device. Future work could look more closely at how the presumed scientific authority of a brainscanning apparatus affects people's willingness to accept specific BCI applications such as passthoughts [**Ali2014a**]. Finally, one subject's skepticism what brainwaves can reveal stemmed from his/her personal medical experiences.

> *(S116) My son has absence seizures, so his brainwaves change.*

This quote highlights how individuals' life experiences might shape the way they engage (or refuse to engage) with brain-sensing devices. In general, this quote and others motivate the need for a rich, qualitative understanding of people's first-hand experiences with brainscanning devices, as well as data collection, in order to understand what role BCI applications such as passthoughts could play in day-to-day life.

3.5 Brain-based authentication as a usecase

TODO Motivating a brain-based probe

1. **TODO** Why brains interesting case given data
2. **TODO** Wonder aloud how to study brain-based case, lay out specs

Brain-based authentiation

1. Authentication

   Authentication seeks to prove that a user is who they claim to be. In computer security, authenticators are classified into three types: knowledge factors (e.g., passwords and PINs), possession factors (e.g., physical tokens, ATM cards), and inherence factors (e.g., fingerprints and other biometrics). Passwords, which rely only on a knowledge factor, offer many benefits [**Bonneau2012**], including their capability for allowing *authorization* (granting use of a system). However, for authentication (verifying an

individual), passwords are easy to compromise, and easy to forget. These shortcomings motivate the overlay of multiple factors of authentication.

2. Multi-factor authentication

   An ongoing problem in authentication lies in balancing strong security (i.e., multiple factors) with usability. As an example, major industry players such as Google and Facebook have strongly encouraged their users to adopt two-factor authentication (2FA), in which a user enters his or her password (a knowledge factor), and subsequently receives a code on their cellphone (a posession factor).

   However, submitting two different authenticators in two separate steps has frustrated wide adoption due to its additional hassle to users. The Apple iPhone, for instance, already supports device unlock using either a user-selected passcode or a fingerprint. The device could very well support a two-step two-factor authentication scheme if desired. However, it is easy to understand why users would balk at having to enter a passcode *and* provide a fingerprint each time they want to unlock their phone.

3. One-step, multi-factor authentication

   Since passthoughts' initial proposal in 2005, more ubiquitous sensing and computing has enabled a number of other strategies for achieving two factors of authentication in a single step. Some work has tested behavioral authentication methods such as keystroke dynamics, or voice. In both cases, the knowledge factor (password or passphrase) and inherence factor (typing rhythm or speaker's voice) are employed [**Monrose1997**]. In contrast, the Nymi band supports one-step two-factor authentication via the inherence factor (cardiac rhythm that is supposed to be unique to each individual) and the possession factor (the wearing of the band on the wrist) [**Nymi**]. More recent attempts have also used gait (from cellphone accelorometers) to perform authentication [**UnifyID2017**].

   However, these existing strategies are susceptible to a variety of attacks. Nymi, for example, does not have a knowledge factor, making it impossible for the user to change the authentication token if the device and biosignal have been compromised. Keystroke dynamics, voice, and gait are all susceptable to "shoulder surfing," in which an attack uses visual or other cues to steal, or improve the chances of guessing, a target's chosen secret. Passthoughts mitigates this attack by nature of the mental gesture. Since the expression of a passthought is not externally visible, the authenticator is impervious to shoulder surfing attacks. Since the thought incorporates a chosen secret (knowledge factor), it can be changed if compromised.

   To assist with the usability and security issues surrounding multi-factor authentication, passthoughts aims to provide two factors of authentication in a single step. A single mental task, or passthought, provides both a knowledge factor (a chosen secret thought) with an inherence factor (the way that thought is expressed for an individual)

[**Chuang2013b**, **Johnson2014**]. Using a custom sensing device, passthoughts could provide an additional posession factor, all in the same step.

The use of EEG as a biometric signal for user authentication has a short history. In 2005, Thorpe et al. motivated and outlined the design of a passthoughts system [**Thorpe2005**]. Since 2002, a number of independent groups have achieved low (less than 1%) false acceptance rates using multi-channel sensors placed on the scalp [**Poulos2002**, **Marcel2007a**, **Palaniappan2008**, **Ashby2011**]. In 2013, one group showed that similar accuracy can also be achieved using a consumer-grade single-channel sensor [**Chuang2013b**]. In particular, the lack of signal diversity from multiple EEG channels can be overcome by allowing the users to choose their own personalized passthoughts (e.g., sing their favorite song in their head). There are two significant consequences of this result. First, the passthoughts approach is no longer constrained by the high cost ($>$ \$10,000 USD) and low usability (gel-based electrodes; aesthetic challenges of an EEG cap) of medical-grade multi-channel devices. Second, because users can choose and easily change their secret mental task, this approach can support one-step two- factor authentication via the simultaneous presentation of the inherence factor (brainwave signatures due to the unique folding structures of the cortex) and the knowledge factor (the secret mental task) [**Chuang2014**].

TODO Brain-based authentication as a probe

1. **TODO** How it meets specs laid out above
2. **TODO** Tease next chapter where we focus on engineers

Chapter 4

Talking to engineers about brain-computer interface

In 2017, both Mark Zuckerberg and Elon Musk announced efforts to build a brain-computer interface (BCI) [**Levy2017**]. One blog post enthusiastically describes Musk's planned BCI as a "wizard hat," which will transform human society by creating a "worldwide supercortex," enabling direct, brain-to-brain communication [**Urban2017**].

A slew of inexpensive brainscanning devices underwrite such utopic visions. 2017 saw a BCI for virtual reality gaming [**Neurable2017**] and brainwave-sensing sunglasses [**Optical2017**] join the already large list of inexpensive, consumer BCIs on the market [**Levy2017**, **Interaxon**, **Grierson2011a**]. These devices, which are typically bundled with software development kits (SDKs), shift the task of building BCIs from the realm of research into the realm of software development. But what will software developers *do* with these devices?

This study employs a technology probe to surface narratives, and anxieties, around consumer BCIs among professional software engineers. We provided a working brain-computer interface to eight software engineers from the San Francisco Bay Area. As brainscanning devices become more accessible to software developers, we look to these BCI "outsiders" as a group likely to participate in the future of brain-computer interface. Specifically, we provided participants with a brain-based authenticator, an application predicated on the notion that a BCI can detect individual aspects of a person, making it a potentially fruitful window into broader beliefs about what BCIs can reveal [**Rose2016a**, **Dumit2004**].

Despite heterogeneous beliefs about the exact nature of the mind, the engineers in our study shared a belief that the mind is physical, and therefore amenable to sensing. In fact, our participants all believed that the mind could and would be "read" or "decoded" by computers. We contribute to an understanding of how engineers' beliefs might foretell the future of brain-controlled interfaces. If systems are to be built that read the mind in any sense, we discuss how such systems may bear on the long-term future of privacy and cybersecurity.

Figure 4.1: A participant uses our brainwave authenticator in his startup's office.

4.1 Brain-computer interfaces & pathways to broader adoption

BCIs allow people to interact with computers without muscular action. Instead, nervous system activity is translated to a discretized (digital) signal. BCIs can be categorized broadly as invasive (requiring implantation) or non-invasive (requiring only external, removable equipment). Non-invasive, consumer BCIs, are lightweight, require minimal setup, and do not require special gels. EEG (electroencephalography) is currently the most viable choice of sensing modality for consumer BCIs [**Carrino2012**].

Historically, researchers have conceived of BCIs as accessibility devices, particularly for individuals with severe muscular disabilities. However, accessibility devices can sometimes provide routes for early adoption, and thus broader use. Speech-to-text, for example, was once a tool for individuals who could not type; eventually, it became adopted as a tool for computer input, now commonplace in IoT devices such as Alexa and Siri. Since accessibility devices can give rise to broader consumer adoption, we ask what such a pathway might look like for brain-computer interfaces. With an expanding array of inexpensive brainscanning

hardware, many of which come bundled with engineer-friendly SDKs, the pathway to a future of consumer BCI increasingly becomes a matter of software engineering.

Thus, we look to software engineers in the San Francisco Bay Area. We use these engineers as a window into broader beliefs about "Silicon Valley," a term we use here to stand in for the technical, economic and political climate that surrounds the contemporary technology industry in the area [**saxenian1996regional**]. While we do not believe only Silicon Valley engineers will influence the future of BCIs, historically, these engineers have a outsized impact on the types of technologies developed for mass consumption, especially with respect to software. As BCI hardware becomes more accessible, and therefore more amenable to experimentation as software, this group once again holds a unique role in devising a consumer future for this biosensor. Indeed, the Muse, and similar devices, have robust SDKs and active developer communities that are building and showcasing BCI applications [**NeurotechX**].

However, we did not want our subjects to have first-hand experience in developing BCIs, as we did not want them to be primed by existing devices' limitations. Instead, we selected individuals who indicated they would be interested in experimenting with consumer BCI devices in their free time. This screening was meant to draw subjects likely to buy consumer devices and develop software for them. We believed that these engineers' professional expertise in software development afford a desirable criticality around our technical artifact.

What brain scans can tell

Brain scanning holds a unique *charisma* [**Ames2015**], not only among researchers in related fields [**Rose2016a**], but among non-experts as well [**Ali2014a**]. Ali et al (2014) found university undergraduates believed a brain scanning device (a fake one, unbeknownst to them) could reveal intimate details of their thoughts, even after receiving a lecture about the limitations of brain scanning technologies [**Ali2014a**]. In that study, participants saw scans of the *brain* as informative with regard to the *mind*, a distinct entity that is potentially more expansive than the brain [**Clark2013**, **Hayles1999a**].

This entanglement of mind and brain has been explored by past work in science and technology studies. For example, Dumit's (2004) study of positron emission tomography (PET) explores utopian (and dystopian) visions of diagnosing mental illness, or even criminality, from scans of a person's brain [**Dumit2004**]. The idea of the mind's "legibility" via computational technologies has been concretely explored by Rose (2016) [**Rose2016a**], who ties together a number of efforts across neuroscience and cognitive science to argue that specific technical implementations from these fields (along with their rhetoric around, and beliefs about the brain) allow *the mind* to be "read" or "decoded."

However, there exists an opportunity to investigate how pervasive such beliefs are among those who are not neuroscience experts, yet nonetheless technical practitioners. Given the recent shift of brain scanning equipment from research tool to consumer electronic device, we ask what software engineers, newly able to develop applications around brain scanning, might build. Answers to this question could have far-reaching consequences, from marketing, to entertainment, to surveillance. In particular, we aim to center how engineers' ideas about

the mind, especially its relationship to the brain and body, inform and constrain their beliefs about what BCIs can (and should) do.

A BCI technology probe

In this study, we use a technology probe to examine the beliefs of software engineers about what BCIs can reveal about the mind. Technology probes are functional apparati intended to both collect data *in situ* from participants, and to inspire participants to reflect on the probes, and on their beliefs more generally [**Hutchinson2003**].

Probes have a long and diverse history within HCI, often referring to a variety of different practices [**Boehner2007**]. In the context of our study, our probe seeks primarily to answer research questions, rather than to figure as one step in an iterative design process. Unlike some probes in past work ours was not intended for longitudinal deployment. Instead, we aimed to gather beliefs about particular technologies and domains through a session of open-ended interaction with a device [**Leahu2014**].

Our probe's unfinished appearance was intended to invite critique and playful experimentation [**Devendorf2016a**, **Leahu2014**]. However, unlike a mock-up or provocation, our probe did function as advertised, allowing participants to interact with the devices in an exploratory and unconstrained way (indeed, many engineers tested that the device's feedback was real). We designed our probe to steer participants away from providing narrow feedback about the interface at hand, and toward sharing their broader beliefs about the brain and mind.

Brain-based authentication

Our study employs a brain-based authenticator as a research probe to elicit engineers' beliefs about BCIs (and the mind and/or brain they purport to sense). This section explains how brain-based authentication works, and why we chose this application for our study.

Authentication (i.e., logging into devices and services) entails a binary classification problem: given some token, the authenticator must decide whether or not the person is who they claim to be. These tokens typically relate to one or more "factors": knowledge (something one knows, e.g. a password), inherence (something one is, such as a fingerprint), or possession (something one has, such as a device) [**Chuang2013b**]. Brain-based authentication relies on signals generated from individual's brains to uniquely authenticate them, which has a number of potential advantages over other authentication strategies (see [**merrill2017future**] for a review). First, brainwaves are more difficult to steal than biometrics fingerprints, which are externally visible, and left in public as one's hands touch objects in the environment. Brainwaves also change over time, making theft even less likely. Second, brain-based authentication requires no external performance, making it impervious to "shoulder-surfing attacks" (e.g., watching someone enter their PIN).

We chose to build a brain-based authenticator for our study for a few reasons. First, having participants use a functioning system helped them imagine how they might use BCIs

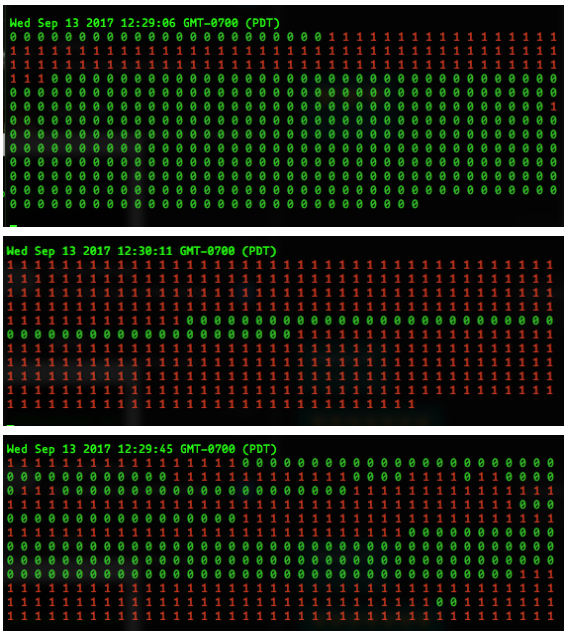

Figure 4.2: Our probe's visualization of 1's and 0's gave our engineers a "raw" view of the authenticator's behavior. Pictured, the UI (a) accepting someone, (b) rejecting someone, or (c) presenting mixed, ambiguous feedback.

themselves. Second, the system is a plausible one, backed by peer reviewed research, thus we expected our participants to judge its claims credible. Third, the system embeds particular assumptions about what brain scanners are able to capture. Our system embeds ideas that our Muse headset can capture aspects of individual brains that are unique; as such, we expect that a working, brain-based authenticator will encourage participants to reflect not only on how a BCI applications might be adopted by the broader public, but also on what BCIs may be able to reveal about the mind and brain, and to critically examine the limits of what BCIs in general are able to do.

4.2 Building the BCI authenticator probe

Implementation

Since we wanted our technology probe to appear portable enough for use in the real world, we decided to use a pre-existing consumer EEG device to build our authenticator. We settled on the Interaxon Muse (Figure 4), a $299 headband that can be worn easily, transmits data wirelessly, and requires no gel to maintain contact between the scalp and electrodes [**Interaxon**]. Using a system that required conductive gel would have signaled to the participants that the technology is still limited to lab settings, and not yet ready for the real world, which could have influenced their responses.

Although the Muse's signal likely contains noise, a perfectly clean signal was not necessary to elicit beliefs from subjects in the context of our technology probe. Further, despite the Muse's small form-factor and dry electrodes, past studies have verified its signal is sufficient quality for some neuroscientific research [**Krigolson2017a**].

Due to the device's battery life and intermittent connectivity when walking, the Muse headband did made a longer-term study impractical. Thus, we opted to perform a study over a short time and in a controlled environment, drawing on past technology probe studies with similar constraints [**Devendorf2016a**, **Isbister2006**].

Data from the Muse was collected via the device's native OSC interface, and stored in a timeseries database. Queries from this database were used to provide training data for a machine learning classifier. In a preprocessing step, we performed a fast Fourier transform (FFT) to generate frequency-domain data from the time-domain data. In the machine learning step, we split a corpus of readings (and labels) into train and validation groups. Using XGBoost [**Chen2016**], we trained a binary classifier on seven different splits of the train group. After the classifier was produced, we validated its performance on the withheld validation set.

Given a target participant to classify, our classifier used any reading from this participant as a positive example, and any reading *not* from this participant as a negative example. Negative examples also included signals with poor quality, and signals from which the device was off-head or disconnected. Ideally, the resulting classifier should produce "authenticate" labels when the device is on the correct person's head, and "do not authenticate" labels at any other time. This classifier could output its labels to a simple user interface (UI), described in the next section.

Interface

As the device produces data, the classifier outputs labels of "accept" or "reject." Our interface displays these labels as a square of 0s and 1s, which filled up as data from the device rolled in (Figure 5.2).

Several considerations motivated this design. First, the UI represents the probabilistic nature of the classification process. Individual signals may be misclassified, but over blocks of time, the classifier should be mostly correct (represented as blocks of mostly 0s by our interface). Thus our simple UI makes visible both the underlying mechanism of binary classification, and its probabilistic nature. Second, because our UI provides potentially ambiguous feedback (as opposed to unambiguous signals of "accept" or "reject"), it allows for potentially richer meaning-making and explanatory work [**Sengers2006a**]. Toward this end, the UI's real-time reactivity ("blocks" of 1s and 0s filled in over time) allows participants to experiment actively with the device, forming and testing hypotheses as to what makes classification succeed or fail.

Finally, our UI gives the probe an "unfinished" appearance. We believed this interface would cause our participants to activate their "professional vision" as tech-workers [**Goodwin1994**], and critique or test the device as if it were a design of their own. Ideally,

we hoped participants would intentionally stress-test the device, or find playful ways of misusing it. These misuses could allow participants to form hypotheses about why and how the device succeeds and fails.

4.3 Methods

We recruited participants by word of mouth. A recruitment email explained that subjects would interact with a working BCI, and be asked their opinions about the device, and about BCI broadly. We screened respondents by their current occupation and stated interest in experimenting with BCIs in their free time.

A total of eight people participated, three of which were women. Participants' ages ranged from 23 to 36. We met with subjects for a single, one-hour session in which we trained and tested a brain-based authenticator, allowing them to interact with it in an open-ended way.

These sessions were designed as a semi-structured interview, interspersed with conversation between the researcher and the participant. Our study protocol was approved by our institutional IRB. Interviews were recorded, and later transcribed. We performed an "issue-focused" analysis of the transcriptions [**weiss1995learning**], allowing topics and themes to emerge during analysis. Subjects names were changed to pseudonyms to protect their anonymity. The remainder of this section describes in detail how subjects interacted with the device during sessions.

Wearing the device

The interviewer began by explaining that participants would wear a BCI, which we would train to work as an authenticator, answering participants' questions about how the device would work. Subjects were told that they would be asked about their opinions on BCIs generally, and that their anonymized voice and EEG data would be collected.

The interviewer asked participants to place the EEG headband themselves, and to assure that the device fits comfortably, at which point the interviewer would begin recording signals from the device. Next, the interviewer would ask participants how they felt about having the EEG device on their head. This question would typically begin a short, open-ended exchange about their past experience with brain-scanning devices, and prior knowledge, if any, of BCIs. This exchange would segue into a broader discussion about the participant's use and relationship with technology, in personal and work life.

After this initial conversation, the interviewer would perform a brief *calibration* step with the participant, in which data are collected to train a custom classifier for use in authentication. Participants would perform a number of tasks, or *mental gestures*, prompted by a stimulus presentation program. These tasks provide a more diverse corpus of an individual's signals, which should enable a more robust (and accurate) classifier. After this calibration procedure, which usually lasted about ten minutes, the interviewer would perform a semi-

structured interview with participants. The interviewer would continue to record data from the Muse throughout this interview.

Using the authenticator

At this point, the interviewer would explain to participants that the data collected thus far would be used to train an custom authenticator for them. The interviewer would explain roughly how the authenticator would work: the probe should *accept* readings when the participant is wearing the device, and *reject* readings in any other case.

Next, the interviewer would run a script that trained our XGBoost classifier (Section 4.2). Participants could watch the training process run, if interested (a few were). After the training process completed, the researcher would set up the UI (Section 4.2) and allow participants to view the classifier's output in real-time using live data from the participant's Muse device. Participants would then see the probe's *accept* or *reject* classifications using the live data from their headset.

After allowing participants to acclimate to the output, and answering any preliminary questions, the interviewer would encourage the participant to experiment with the authenticator, and share any impressions, reactions or ideas. The open-endedness of this session was meant to encourage participants to explore the device's capabilities and limitations, free of particular tasks to accomplish. However, we suspected that our participant population would be particularly prone to "hypothesis-testing," exploring the devices limitations by building theories about how it might work. We structured the session around this assumption, preparing to ask participants to think aloud as they explored the device's capabilities.

After some free-form exploration (usually involving some back-and-forth with the participant), the interviewer would transition into a semi-structured interview, which would occur with the device still active. The interviewer would ask participants to unpack their experience, and lead them to explore what they felt the device could reveal about them. After some discussion, the formal interview would conclude, and the participants would remove the Muse device from their head.

4.4 Experiencing the authenticator

In general, we found particular reflections to come at different points in the interview protocol. Critiques (and questions) about the device narrowly tended to come as soon as engineers placed the device on their heads. Reflections on the BCI broadly, and its future trajectories, tended to come after viewing the probe's feedback for some time. As these conversations progressed, participants naturally tended to reflect on what future BCIs could do. Subjects would typically relate the capacities of the probe, and of possible future technologies, to their ideas about the mind, body or brain. The probe continued to run during these discussions. Toward the end of the interview, the researcher would prompt participants to reflect on any anxieties they might have about the future of BCIs (interestingly, only one participant raised

this subject on their own). The remainder of this section is organized to roughly mirror this common order of participants' reflections during interviews.

Using the BCI probe

Our working authenticator elicited diverse reactions from the engineers in our study. Almost all participants cracked jokes after putting on the headband (three subjects commented that they felt like they were "from Star Trek"). All participants except Joanna said they would not wear the device in public, though a few conceded that they might if the headsets were more common. Terrance commented, "If most people are doing it, then it's fine. Sort of like stock speculation."

Perceptions of the authenticator's accuracy were mixed. Four participants found that the authenticator worked well for them. For these participants, the authenticator consistently rejected blocks when the headset was off of their head, or worn by the researcher.

On the other hand, four participants found the probe consistently rejected every reading, whether it came from them or the researcher (i.e., they experienced false rejections, but not false acceptances). These subjects often tried to remedy the situation by attempting tasks they had rehearsed, typically with mixed success. Most of these subjects concluded that there was not enough training data to produce reliable classification, but that such a system would work with a larger corpus. In contrast, Alex, a 30 year-old founder of an indoor agriculture startup, blamed himself, saying "I must not produce very distinguishable thoughts."

Those participants who felt the probe's authentication was reliable tended to center their explanations on why it worked. Participants who experienced less consistent accuracy with the authenticator tended to center their explanations on how the device might be improved, e.g. with better or more comprehensive sources of data. This impulse to "fix" likely speaks to our participants' general tendency to engineer working systems, which extended in our case even to this experimental technology.

As we hoped, the engineers engaged critically with the technical implementation of the probe. In general, engineers asked about the machine learning infrastructure underlying the authenticator, and several participants (particularly John, Mary and Alex) asked specific questions, and made specific recommendations, diagnosing issues with the authenticator by thinking about the diversity and size of the training set. Almost all participants noted the authenticator worked better when they were not looking at the visual feedback from the user interface. Participants generally theorized that this might occur because they were not viewing feedback when training the classifier. In these cases, the engineers appeared to apply their domain knowledge to their observations in using our technology probe.

Reflecting on the future of BCI

Our technology probe caused almost all of our participants to speculate on the future of BCIs generally. To most participants, the future of BCIs seemed to be largely pre-determined.

One of our participants, Terrance (a 24 year-old software engineer at a small transportation startup), removed the headband to inspect it, and commented on its awkward visibility. In doing so, he reflected on the future of BCIs, speaking in no uncertain terms about a future of computer-mediated "telepathy."

> Things just get progressively smaller until they disappear. And one day this'll just be an implant in my brain, doing crazy things. It'll be interesting socially, how people come to terms with it, when it's just an implant, or at least very pervasive ... I could send you a message, and it could be like you're thinking it yourself, even if you're on the other side of the Bay. *(Terrance)*

Terrance believed that BCI *will* become more prevalent: not just that smaller sensors will lead to more effective or usable BCIs, but that they will also result in greater uptake of the technology. While he references the social dimension of their adoption, he indicates that people will need to "come to terms with" the developments, rather than providing direct agency to users who may choose to adopt the technology or not.

Two participants felt less sure that such a future of pervasive BCI would ever come to pass. Elizabeth, a 30 year-old front-end engineer, noted skepticism about signal quality, or usefulness outside of persons with disabilities. Mary, a 27 year-old software engineer at a large company, pointed to social reasons for her skepticism. In reflecting on the relative accuracy of the probe's authentication performance during her session, she commented that "90 plus percent" of people would be "totally freaked out" by brain-computer interfaces generally. She continued to say that companies may themselves stop BCIs from becoming too pervasive or advanced.

> I feel like those companies, even if this were feasible, there's a moral quandary they philosophically have not figured out. They will not let the research get that advanced ... I just don't imagine them being like, "okay computer, now read our brains." *(Mary)*

While the probe was effective in spurring subjects to talk about issues around BCIs, its accuracy as an authentication device did not seem to alter participants' belief in BCI's future as a widespread technology. Unsurprisingly, the four subjects who experienced reliable authenticator accuracy all expressed that BCIs would become commonplace in the future. However, only Joanna connected the device's poor performance in her session with a probability of ongoing accuracy issues for BCIs in the future. The other three subjects who felt the device did not perform accurately all offered explanations as to why, and explained that future devices would fix these issues.

Mind, brain, body

During their interactions with the probe, almost all of our subjects discussed their deeper beliefs about the nature of the mind, and its relationship to the brain and body. Since

participants discussed the future trajectory of BCIs led to discussions while the probe continued to work (or fail), the subject often arose of what BCIs might be able to detect, even theoretically. As one example, John, a 26 year-old software engineer at a small chat startup, noticed that the authenticator only worked when he was speaking, but not when he was listening to the researcher. He offered an explanation for the discrepancy.

> There's probably some kind of fundamental difference between creating thoughts and consuming thoughts. You're still making thoughts, right, but it's almost like programming versus being programmed. *(John)*

When pressed on how strictly he meant his metaphor of programming, John confirmed that he meant it quite literally, saying, /"I think we are just computers that are way more sophisticated than anything we understand right now."/ We return to this strictly computational account of the mind as "just" a computer in the discussion.

Mary gave a computational account of mind that was more metaphorical than John's, drawing on comparisons between machine learning and the mind. She cited the many "hidden layers" in deep neural networks, and that, like in the brain, "information is largely distributed." While she believed deep learning models and the brain were "different systems foundationally," she said "there are patterns" that relate the two to one another, and indicated that advances in deep learning would spur a greater understanding of the brain.

Although six of our participants provided a largely computational account of mind-as-brain, not all did. Joanna, a 31 year-old engineer who previously completed a PhD in neuroscience, felt that the mind was "the part of the brain I am aware of, the part that is conscious." She believed that neurotransmitters throughout the body have a causal relationship to what happens in the mind, but do not constitute the mind themselves; the contents of mind occur physically in the brain, and the brain alone. In other words, her account is one of "mind as conscious awareness," and while unconscious phenomena affect mind (e.g. the body, environment), they are not part of the mind *per se*. Interestingly, the probe did not work well for Joanna, and she felt confident that its poor performance was due to contaminating signal from her body (a theory she tested, and validated, by moving around and observing the probe's feedback).

Meanwhile, in one subject's account, the mind extended beyond the confines of the body. Terrance felt that there was "no meaningful difference" between the body and brain, nor between the body and the physical environment at large, saying that "you can't have one without the other." He believed that all three of these entities constitute the mind in a mutually-dependent way. However, Terrance indicated that the mind is still strictly physical, as are these three entities. Although Terrance did not provide details on how exactly the mind extended beyond the body, it is interesting to note this position's similarities to Clark's (2013) account of the extended mind [**Clark2013**], or Edward Hutchins's work on distributed cognition [**Hutchins2005**], though Terrance was familiar with neither.

Participants also offered differing levels of confidence in their beliefs about the nature of the mind. Joanna (who has a background in neuroscience) reported that "we do not know

everything we need to know" about how the mind works. Three other subjects reported similar beliefs. However, those subjects with a computational account of mind tended to feel more confident that their account was substantially accurate.

> I think the consensus is that the body is mostly like the I/O of the brain. *(John)*

John's account here implies that a sufficiently high-resolution brain sensor would accurately capture all of a person's experiences. John confirmed this explicitly, saying *"if you could 3D print a brain, and apply the correct electrical impulses, you could create a person in a jar."* In this computational metaphor of I/O (input/output), the body itself does not have agency; instead, the body actuates the brain's commands (output), and senses the environment, sending data to brain for processing (input).

Reading the mind

As discussed in the previous section, every participant's account of mind was strictly physical, rooted mostly in the brain, in a few cases in the body, and in one case extending beyond the body to the physical world. With this physical understanding of the mind, it is not overly surprising that all participants believed it would someday be possible for a computer to read or decode the contents of the human mind. No participants expressed hesitation when asked about such a proposition.

For example, Alex did not feel comfortable providing a specific physical locus for the mind. Although he did not feel the probe was accurate for him, he took great pains to express his belief that such a device could work, though not necessarily by sensing the brain.

> We're driven by single-celled organisms in ways we don't really yet understand, but... there's got to be some sort of physical storage of memories or experiences. We just haven't quite learned how to read it yet. *(Alex)*

Though it leaves open room for a variety of interpretations about the exact nature of mind, Alex's view is explicit that thoughts are physical, therefore *can* be read, and *will* be read with some future technology.

There was a great deal of heterogeneity in the way this belief was bracketed or qualified. Joanna felt that there would "always be parts of the mind that can't be seen." She likened the question to the way that other people can know some parts of another person's mind, e.g. through empathy; their perspective, however, would always be partial, and she felt the same would be true for machines.

However, some participants did not bracket their belief that machines would someday read the mind. Participants for whom the authenticator worked reliably typically said that a mind-reading machine was "absolutely possible" (Mary) or "just a matter of the right data" (Alex). Participants who did not feel the authenticator was accurate described current state-of-the-art as "crude" (John) or "low-granularity" (Elizabeth).

Even Terrance, who believed the mind extended beyond the confines of the body, felt that the mind was readable by machine. After he stated his personal belief in a mind that extended to the physical environment, the experimenter asked what consequence this belief might have for the future of BCIs.

> Practically, it has no implication. We could still devise an authentication tool that does the job, and it doesn't matter. Maybe in some way there could be this ESP thing where you could somehow read my thoughts... If we want to do something, we will find a way. *(Terrance)*

Terrance's language here belies broader narratives of positive technological progress (notions of "[moving] forward," and that "we will find a way"). Despite his personal beliefs about the "true" nature of the mind, he felt that engineers would manage to build the systems they intended to build, even ones with a much higher specificity than those available today (e.g. an "ESP device").

BCIs for everyone?

Generally, participants stated (implicitly or explicitly) that BCI technologies would become smaller, less expensive, more accurate, and therefore become prevalent as a consumer device. Only Mary raised the question of how institutions exert agency over the artifacts they create. Where most subjects indicated BCIs become smaller and thus more pervasive, Mary indicated that companies have beliefs, which affect what devices and technologies they produce. Specifically, Mary spoke of a "quandary" between advancing technology on one hand, and systems' autonomy on the other. She viewed this reluctance to allow systems to become more autonomous as a signal that certain technologies, potentially including BCIs, may *not* be developed for ethical, moral or philosophical reasons.

Interestingly, the other seven engineers in our study expected a future in which BCIs are pervasive, in spite of their unwillingness to wear our probe's headband in public. Some subjects believed the device's awkward, outward visibility might be mitigated by future miniaturization. Other subjects felt that social norms may simply change if the device became pervasive. This latter attitude is reminiscent of those around Google Glass, which shared an awkward (and, in practice, often stigmatizing) visibility [**wong2016product**]. Future work might draw out the relationship of Google Glass's imagined future to that of BCI, perhaps as a way of learning lessons about possible commercial failures, and how engineering communities may have failed to foresee them.

BCI anxieties

An important counterpoint to emerging technologies is the anxiety that rises along with them [**Pierce2017**]. Interestingly, engineers in our study expressed no strong anxieties regarding the development of BCIs, for the most part. Regardless of their experiences with

our probe, participants felt that BCIs would be developed, and would improve people's lives. Participants mentioned domains such as work, safety, and increased convenience in the home.

Only Mary reported existential anxiety about the possibility of machines that could read the human mind. She reported a technology to be "absolutely possible," and referenced the probe's continuing high accuracy as we spoke. However, in stark contrast to Terrance, Mary *feared* such a development would occur sooner rather than later.

> I hope it's fifteen years out, but realistically, it's probably more like ten. *(Mary)*

Despite Mary's prior statement about the power of institutions to change the course of technical developments, here she seems to indicate that such course changes will not occur, or that they will converge on machines that can read the mind.

When pressed on downsides, the participants who did not volunteer any anxieties about BCI initially did mention security (especially the "leaking" of "thoughts") as a concern. For example, Elizabeth did not report any particular anxieties about BCIs in general, "if the proper protections are in place." Pressed on what those protections might look like, she cited encryption as a solution to privacy concerns. Terrance, who expressed wanting BCIs to become more widespread, described in deterministic terms the cybersecurity issues such devices might pose.

> If there are security holes - which there almost certainly will be - then what happens when I'm leaking my thoughts to someone? What if I'm thinking about the seed phrase for my Bitcoin wallet... and then you put it in this anonymized dataset ... and I lose all my coins? What then? *(Terrance)*

Even alongside his concern, Terrance very much wanted a mind-reading machine to exist. He mentioned a desire for a programming assistant that would somehow speed up the process of software development. Since Terrance's conception of BCI presents high stakes with regards to privacy and security (he variously mentioned "telepathy," and an "ESP device," implying a high degree of specificity with regard to what BCIs can resolve), it is telling that he thought primarily of using BCIs to become a more efficient engineer, rather than concerns around privacy or potential harm. Later in the discussion, we unpack further how larger cultural tendencies in Silicon Valley might shape the way engineers build BCI systems.

4.5 Discussion

We find that engineers hold diverse beliefs about what the mind is, what the brain is, and about the relationship between these entities. However, all of these engineers shared a core belief that the mind is a physical entity, one that machines can and will decode given the proper equipment and algorithms (Section 6.1). Despite this belief, engineers did not largely express concerns about privacy or security (Section 6.2). As BCI startups continue to grow, we propose further work within technical communities, with a sensitivity toward emerging

narratives, so that we may instill criticality among this emerging technical practice (Section 6.3). We conclude with avenues for future work focusing on different communities of technical practice (Section 6.4).

Physical mind, readable mind

Although our engineers broadly believed BCIs would become pervasive as consumer devices, we found no consistent visions of what such a future might look like. Instead, and to our surprise, we found a shared belief that there exists a physical mind that can be "read" or "decoded" by machines, despite participants' heterogeneous beliefs about its exact nature. Interestingly, only one participant shared any anxiety about this prospect with the researchers; the other participants reported looking forward to such a possibility.

Crucial to beliefs about the machine-readable mind were frames of the mind as physical, and therefore amenable to sensing. In many cases, subjects would use analogies to computation in making this point. For example, John observed an anomaly in the authenticator's performance (it did not work when he was listening to the experimenter speak). He theorized that the states are distinguishable, because speaking "is like programming" and listening to someone speak "is like being programmed". In this case, John's observations about the BCI met with his pre-existing notions of the mind, producing a hypothesis for what "brain states" might exist *and* what states Muse headset might be able to detect. Hypotheses such as these could be consequential, as they might provide ideas or starting points for engineers looking to build systems. Our results highlight the importance of both pre-existing beliefs and particular interactions with BCIs in structuring engineers' understandings.

Broadly, engineers' beliefs about the mind-as-computer metaphor (Section 4.4) could provide starting points for engineers to build BCIs in the future. This computational view of mind has been popular among engineers at least since the "good old-fashioned AI" (GOFAI) of the 1950s. While much work has critiqued this stance from various angles [**Agre1997**, **Hayles1999a**], those same critiques have acknowledged the role these metaphors have played in the development of novel technologies: If the mind is a machine, then those tools used to understand machines can also be used to understand the mind. Here, we see this metaphor return, its discursive work now focused on biosensing rather than on artificial intelligence. Of course, these metaphors illuminate certain possibilities while occluding others [**Hayles1999a**]. As such, future work should follow past research [**Agre1997**] in understanding what work this metaphor might do in its new domain of computational mind-reading.

Even those participants who did not subscribe to computational theories of mind still believed the mind to be strictly physical. These subjects all agreed that computers could someday read the mind, precisely because of its physical nature. While our results indicate that engineers believe the mind to be machine-readable, some work indicates that non-engineers may share this as well [**Ali2014a**]. Future work could further investigate this claim more deeply in the context of consumer BCIs. If so, a machine designed by engineers and purported to read the mind might find acceptance among a broader public audience.

Those subjects with a computational account of mind tended to feel more confident that their account was substantially accurate. John referenced "the consensus" in justifying his beliefs about the mind being equivalent to the brain. It is worth asking whose consensus this might be: that of neuroscientists, philosophers of mind, cognitive scientists, or engineers? In any of these cases, engineers' confidence in their beliefs could have implications for what types of systems are considered buildable, and where engineers might look to validate their implementations. As products come to market, professionals in the tech industry must find ways of claiming their devices to be legitimate, or working, to the public (consumers), to potential investors, and to other engineers. These claims of legitimacy could prove to be a fruitful window for understanding the general sensemaking process around these devices as their (perceived) capabilities inevitably evolve and grow alongside changing technologies.

A future for privacy and security

Since the engineers in our study believed the mind to be readable, an important question remains around the consequences for the future of consumer privacy and security. Our participants largely acknowledged that "leaking" thoughts through security holes was a valid concern, and one participant claimed that these exploitable holes will "almost certainly" exist. However, the types of threats that engineers referenced may not square with the notion of BCIs as a device for the masses. For example, Terrance's concern about someone stealing his Bitcoins through some BCI-based attack involves a technology which for now remains niche. This imagined scenario demonstrates how the security (and privacy) concerns of engineers may not match that of the general public. Such mismatches could have consequences for the types of systems that are designed, and whose needs these systems will account for.

Crucially, discussions about privacy and security concerns did not cause any participants to reflect further on the consequences of pervasive BCIs, nor did they deter enthusiasm for the development of these devices. These findings indicate either that engineers are not be inclined to prioritize security in the systems they build, or that they have resigned themselves to the inevitability of security holes in software. In either case, our findings suggest a long-term direction for cybersecurity concerns. These devices carry potentially serious security and privacy consequences. If our engineers will try to build devices that make judgments about the inner workings of a person's mind, future work must critically examine how to protect such systems, and the people who use them.

Implications for the design of mind-reading machines

Our findings do not indicate a singular path for the future of BCIs. Instead, they indicate an undercurrent of belief among Silicon Valley engineers in the possibility of technologies that can read the contents of the human mind. Crucially, our study revealed narratives not just around BCIs, but around the nature of the brain and mind generally, which in turn legitimize narratives about the possibility of mind-reading machines.

Despite these beliefs about what BCIs are capable of, only one participant in our study reported that ethical issues around privacy or security might deter their development. We hope engineers will become more reflexive about these beliefs around BCI, and more critical about their downstream potential for harm (e.g. surveillance). Much as utopian dialogues around the potential of the World Wide Web missed risks to privacy and security, so might similarly utopian ideals of mind-reading machines.

Since the engineers in our study believed BCIs could perform this potentially invasive "mind-reading," why did they largely want such BCIs to be built? Explanations might be found by relating the narratives we uncover to existing social and economic value systems within Silicon Valley communities. Biohacking, for one example, has become an established part of Silicon Valley culture, through dieting (e.g. Soylent, fasting), or more extreme forms of body modification (e.g. chipping) [**Dolejsova2017**]. Underlying all of these cultures is a mechanical model of the body, which facilitates notions of optimization and experimentation. How might BCIs (especially ones that purport to read thoughts) work their way into these already-established cultural patterns? We note that existing consumer BCIs already situate themselves in this context: the Muse headset we used in this study markets itself primarily as a meditation trainer (its advertising copy claims to "remove the uncertainty from meditation") [**Interaxon**]. Examining how BCIs perform discursive work in engineering communities will allow us to better understand engineers' intents as these devices begin to emerge, and help us trace these intents forward as devices are re-imagined, remixed and repackaged for other groups of users in the future.

In the nascent field of consumer BCI, researchers and designers should remain in touch with the beliefs of engineers. We pinpoint beliefs about the mind, and its readability by emerging biosensing devices, as especially an critical facet. Doing so will allow design to remain preemptive rather than reactive as software for consumer BCI emerges. Designers and researchers must not remain on the sidelines; as devices come to market, we must become actively engaged in engineers' beliefs (and practices). These systems hold the potential for exploiting an unprecedented level of personal data, and therefore present an high potential for harm. As such, the area presents a new locus for researchers and designers to engage critically with technical developments.

Future work

Software engineers are a diverse group, and the geographic confines of Silicon Valley do not describe all communities worldwide. Future work could explore communities in different places. Engineers in non-Western contexts may hold different cultural beliefs about the mind, which could lead to vastly different findings.

Professionals who work in machine learning could present another participant pool for future work. Machine learning is a critical component of BCIs, and many contemporary techniques, particularly deep learning, use neural metaphors to interpret and designing algorithms [**ba2016using**]. Thus, practitioners of these techniques may be inclined to draw

metaphors between the brain and the algorithms they employ, which could color their understanding how and why BCIs work or fail.

Future work could allow participants to take an active, participatory role in the analysis of their data, and/or in the design of the BCI system. Although our participants had the technical expertise required to perform data analysis and systems engineering themselves, we did not have participants do any such analysis for this study. This participatory approach will also help us expand our understanding from engineers' beliefs to engineers' practices, as they relate to the emerging domain of consumer brain-computer interfaces. Participants might form their own interpretations of what the data mean (or can mean), building understandings that could differ from those we observed in this study.

4.6 Conclusion

As engineers in the San Francisco Bay Area, the participants in our study sit at an historical site of techno/political power. Our technology probe indicates these engineers believe the mind is physical, and therefore amenable to sensing. What are the consequences for the rest of us? We hope our study will encourage engineers to closely examine the potential of these devices for social harm, and encourage researchers to remain closely attuned to this emerging class of consumer biosensor.

Chapter 5

Who are you really? Probing engineers on authentication and the ground truth of identity

Just before Christmas in 2017, a 6-year old child found her mother sleeping on the couch. She gently picked up her mother's thumb, placed it on the fingerprint scanner of her iPhone, and ordered herself $250 worth of Pokemon presents [**Johnson2017**].

In response to long-standing issues with passwords [**Selyukh2017**, **Palma2017**], consumer devices are increasingly turning away from passwords and toward *inherence-factor authentication*, which rely on properties of the body, such as fingerprints or facial scans. These systems are diverse in their sensing modalities, but similar in their use of digitized identifiers from the body as tokens of personhood. Whereas passwords can be given to others, these identifiers by design cannot be distributed: they are meant to verify not access, buy particular bodies.

However, inherence-factor authentication brings its own challenges: Apple's FaceID cannot distinguish between twins [**Hern2017**]; a child can use her sleeping mother's finger to authenticate [**Johnson2017**]. Such shortcomings beg the question: what is the ground truth for the personhood inherence-based classifiers seek to verify? How does this personhood relate to the bodily identifiers they collect?

In this study, we built a system that uses a brainscanner to authenticate users based on their neural signals. This system was highly experimental, prone to false acceptances and false rejections. It produced a continuous, probabilistic, and often ambiguous "determination" about the identity of the person wearing it.

We provided this system to software engineers in the San Francisco Bay Area, aiming to use this probe to surface their beliefs about the "true" nature of the self, and its relationship to biometric identifiers generally. These engineers represent a community of technical practioners who, through their position in a global software industry [**saxenian1996regional**], have the potential to actively construct notions of selves through authentication or other applications. By surfacing their ideas about personhood, we hope to raise alternatives for

design that are sensitive to various conceptions of selfhood.

Although we set out expecting to discover how engineers' beliefs about the self mirror the notions of self embedded in typical authentication systems, we instead find that these software engineers hold complex notion of self involving multiple, material contingencies over time, which existing authentication systems do not capture. Through these diverse notions of personhood, we possibilities for the design of authentication systems, and other systems in which the self is modeled (explicitly or implicitly). distill implications further when crystalized.

5.1 Background

Authentication, bodies persons

This section traces how passwords' shortcomings motivated a shift to inherence-based authentication, and how this shift presented the novel challenges addressed in the present work. This section concludes by claiming that, in their quest to verify personhood, inherence-based authentication systems make particular epistemological commitments about how persons are defined. These persons or "selves" are what systems are tasked with classifying; they are the ground truth through which a system's success and failure are understood.

1. A brief history of authentication

   In computer security, authentication schemes seek to verify that a user is who they claim to be. Authentication is traditionally established via a taxonomy of "factors": something one knows (knowledge, e.g. a password), something one is (inherence, e.g. a fingerprint), or something one has (possession, e.g. a keycard). For enhanced security, "multi-factor" authentication overlays multiple of these properties. A common two-factor (2FA) scheme requires the knowledge factor of one's password along with the possession factor of one's phone.

   The concept of using a secret phrase for authentication has a largely military history, insofar as its use has been recorded. US soldiers shared a famous challenge-response password on D-Day: were two unfamiliar soldiers to meet, the first would say *"flash,"* the second would reply *"thunder,"* and the first would confirm, *"welcome"* [**lewis2004d**]. However, the use of passwords was documented as early as 146 BC among the Roman military by the Greek historian Polybius [**mcging2010polybius**] (6.34).

   Passwords have been used to manage access in computer systems since at least 1961 [**crisman1965compatible**]. They offer many benefits in a computing context: they require no special hardware to enter, beyond a keyboard, and they can be shared easily between people. This latter feature makes passwords a tool for *authorization* (of access) rather than a tool for *authentication* (of personhood).

2. A shift to inherence

Despite passwords' strengths, high-profile leaks and breaches serve as regular reminders that passwords have serious security issues, particularly around usability [**Bonneau2012**]. In response to the shortcoming of passwords, the Apple iPhone introduced inherence-based authentication in 2013. This fingerprint scanner stems from a long forensic history of fingerprint analysis (and much crticism about the practice's accuracy) [**nelson2010america**]. At the time of its introduction in computer security, the paradigm were seen as less prone to "security mistakes" on the part of users. Based on properties of the body, they are widely seen as more difficult to steal than passwords, and particularly less suceptible to attacks involving remote access.

However, the integration of biometrics creates a new problem for the systems that utilize them: their access is now predicated on characteristics of bodies. Rather than establishing knowledge of a token, inherence-based systems establish *personhood* using physical tokens from the body [**nelson2010america**]. In doing so, these systems actively construct a notion of the self around the tokens they are able to detect.

In this study, we ask: How do these constructed selves match to beliefs about the *true* ground self of personhood? As we discuss in the next section, we look to software engineers in particular, whose beliefs we consider particularly relevant in reflecting on what types of authentication systems have already been deployed (are their models sufficient?), and into surfacing alternative systems that may reflect other conceptions of self or personhood.

The beliefs of software engineers

Our focus on software engineers, particularly in the San Francisco Bay Area, is motivated by this group's outsized impact on the sorts of systems that are developed and deployed globally. We believe this group's beliefs about the nature of the self will have a tight relationship to their technical practice. By surfacing their beliefs, we will be able to better anticipate the systems they may build, and also to help raise alternatives for design that are compatible with engineers' beliefs.

What we do not yet know is how engineers' beliefs about the self inform the design of authentication systems generally. Following Agre's study of artificial intelligence practitioners, [**Agre1997**], we note that it is also unclear how philosophically committed systems designers are to the epistemological implications of authentication systems. Would they agree what current authentication systems offer practical workaround for pragmatic ends? Or would they argue that there is one true self per person, and/or that existing systems correctly identify the ground truth of personhood?

Technology probe

This paper seeks to grapple with these questions by examining what this group believes the self *is*, and how current authentication systems map to their beliefs. In doing so, we hope to

Figure 5.1: The InteraXon Muse. The headband contains flexible, electromagnetic sensors worn over the forehead, and conductive rubber electrodes worn over the ears.

find room both to critique the commitments underpinning authentication systems, to raise alternative systems for authenticating persons, and to think about their consequences. We hope this investigation can move toward a role for research-through-design in cybersecurity, a point we will return to this point in the discussion.

In this study, we presented participants with a technology probe. Technology probes are functional apparati intended to both collect data *in situ* from participants, and to inspire participants to reflect on the probes, and on their beliefs more generally [**Hutchinson2003**]. Probes have a long and diverse history within HCI, often to a variety of different practices [**Boehner2007**]. In the context of our study, our probe seeks primarily to answer research questions, rather than to act as a step in an iterative design process. Unlike some probes in past work, ours was not intended for longitudinal deployment. Instead, we aimed to gather beliefs about particular technologies and domains through a session of open-ended interaction with a device [**Leahu2014**]. TODO something about short study duration / precedent for short-duration probes

Brain-based authentication

As our technology probe, we focused on brain-based authentication as a specific application to bring to our participants. Brain-based authentication uses neural signals, detected from a brainscanning device, as a biometric to authenticate users. This section explains why we picked brain-based authentication as an application area for our technology probe.

Brain scanning holds a unique *charisma* [**Ames2015**], not only among researchers in related fields [**Rose2016a**], but among non-experts as well [**Ali2014a**]. A particularly pervasive facet of brainscanning's charisma is a reductive neuroessentialism [**Vries2007**], in which the *brain* becomes equated with the *self*. In his study of PET scans, Dumit outlines how the rhetoric around this scanning methodology spurred utopian (and dystopian) visions of diagnosing mental illness, or even criminality, from scans of a person's brain [**Dumit2004**]. Dumit raises the question of how notions of *personhood* become entangled with the brain through the technical and social affordances of PET's imagery.

Because of this pre-existing entanglement between the brain and the self in technoscientific discourse, we decided to focus our probe around the same topic. In the paradigm of brain-based authentication, patterns of brainwaves, however dynamic, become equated with the selves who produce them, using the statistical infrastructures of machine learning algorithms. In so doing, the system implies an equiveilance between the brain (as a physical organ) and the self: From the perspective of brain-based authentication systems, a person *is* the way their brain works. Further, while the brain-based authentication paradigm is explored by much academic research [**Chuang2013b**, **Marcel2007a**, **Maiorana2016**, **Marcel2007a**, **Thorpe2005**], it remains both "sci-fi" enough to trigger the imagination of participants, it is also vague enough to produce skepticism and concern as to whether the device will be consistently accurate.

While past work has looked at academic visions around selfhood and the brain [**Rose2016a**, **Sample2016**], less is known about how other technical practioners may understand these

notions, or their relation to one another. As research into brain-based authentication continues [**merrill2017future**], re-examining these findings from neuroscience and STS among software engineers may also shed light on novel considerations for design.

We decided to use the InteraXon Muse (Figure 5.1) as our physical sensing device. The Muse is an EEG-sensing headband containing four electrodes, sold commercially for 249USD. Although the Muse's signal can be contaminated by noise from muscular movement or environmental radiation, its signal has been robust enough for use in some neuroscience studies [**Krigolson2017**, **Karydis2015**], In our own tests, we found the Muse produced a signal clean enough to provide robust classification accuracy for the authors. Further, the device is light and portable, and does not require special gels, making it minimally invasive for our purposes in designing a probe.

1. Designing the probe: Emulating the debugging experience

   While the brain-based authenticator probe had a natural thematic relationship to our research questions, we designed the specific technical artifact to raise the intended questions (and criticisms) among our participants as they interacted with it. Software engineers design systems that rest on rule-based reasoning, yet attuned to the "edge-cases" that arise from the messiness of human realities. Thus, we believe this group would be well-attuned to the limitations and needs of a technical system design. Although these engineers may be "caught" by aspects of the charisma surrounding brain-based systems, they are also well-equipped to disagnose issue, and surface the consequences of these systems. In the hopes of triggering our participants' "professional vision" [**Goodwin1994**] as software engineers, we designed our probe to mimic the experience of sitting at one's terminal and debugging (Figure 5.2).

   Our probe's unfinished appearance was intended to invite critique and playful experimentation. However, unlike a mock-up or provocation, our probe did function as advertised, allowing participants to interact with the devices in an exploratory and unconstrained way (indeed, many engineers tested that the device's feedback was real). We hoped the live nature of the feedback would encourage subjects to discuss the probabilistic judgments of the system's machine learning classifier, a technical configuration that has become the mainstay of modern inherence-based authentication.

5.2 Methods

Participants

Our subjects were recruited from Silicon Valley software engineering startups. Using personal connections and mailing lists, we screen specifically for subjects We used snowball sampling to reach additional participants who met these criteria. Historically, these engineers have a outsized impact on the types of software developed for mass consumer use [**saxenian1996regional**]. As sensor hardware becomes more commercially accessible, and

therefore more amenable to experimentation at the software, this group once again holds a unique role in devising a consumer future for this biosensor. Due to the pervasive culture of unstructured experimentation or "hacking" within Silicon Valley technical culture [**saxenian1996regional**], we screened for participants who expressed an interest in developing BCI equipment as a hobby. We assumed this group would be more invested in analyzing and critiquing the technology probe.

A total of eight people participated, three of which were women. Participants' ages ranged from 23 to 36. We met with subjects for a single, one-hour session in which we trained and tested a brain-based authenticator, allowing them to interact with it in an open-ended way.

Study protocol

Table 5.1: Mental gestures used in calibration section. Each gesture was performed ten times, for ten seconds each. The authenticator's classifier was trained with both these recordings, and unlabeled recordings from the semi-structured interview.

Task	Description
Breathe	Breathe deeply with eyes closed
Word	Choose a word, repeat it mentally
Phrase	Choose a phrase, repeat it mentally
Face	Choose a person, imagine their face
Sport	Imagine performing a sport-related action
Song	Choose a song, imagine hearing it

With the subjects' permission, we made audio recordings of our sessions with participants. Our study protocol was approved by our institutional IRB. Interviews were recorded, and later transcribed. We performed an "issue-focused" analysis of the transcriptions [**weiss1995learning**], allowing topics and themes to emerge during analysis. Subjects names were changed to pseudonyms to protect their anonymity.

First, participants would wear the device, become acquainted to its fit and appearance, and confirm the device is working. Next, the experimenter would begin a calibration session, in which the subjects would be prompted to choose mental gestures (Table 5.1), and perform them several times. After this calibration, the experimenter would ask participants to review which mental gestures they chose. Following this discussion, the interviewer would ask subjects what their favorite and least favorite tasks were.

After some discussion on why subjects preferred certain tasks over others, the interviewer would ask, "Which task do you think would be most unique to you?" This question, which typically resulted in several follow-ups, was aimed at learning what subjects believe the device can reveal. The interviewer would attempt to drill as deeply as possible (without leading subjects) to discuss the particular mechanisms they believe to be at play. As an example, the researcher would first ask subjects which tasks they felt might be more unique

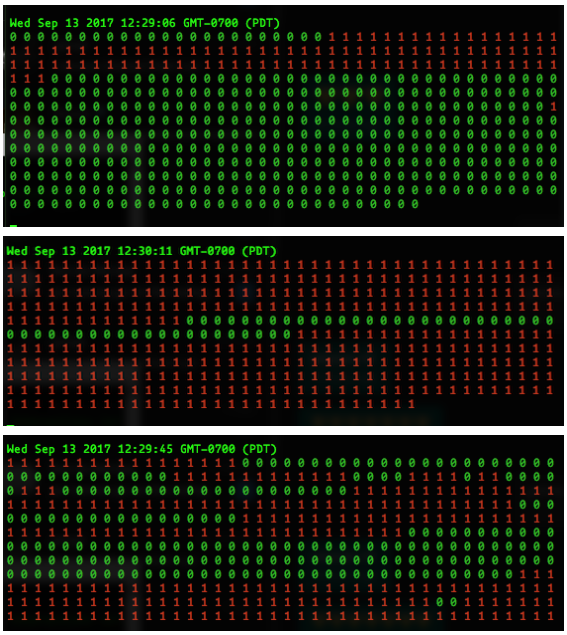

Figure 5.2: Our probe's visualization of 1's and 0's gave our engineers a "raw" view of the authenticator's behavior. Pictured, the UI (a) accepting someone, (b) rejecting someone, or (c) presenting mixed, ambiguous feedback.

to them. In probing why subjects felt this way, the researcher would encourage subjects to reflect on what the device might really be sensing, and what sort of "resolution" it might have. From there, the researcher might ask what any device (even theoretical ones) might be able to detect.

At this point, the interviewer would train the authenticator's XGBoost **[Chen2016]** classifier on samples collected from the subjects. The interviewer would explain to participants roughly how the authenticator would work: the probe should *accept* readings when the participant is wearing the device, and *reject* readings in any other case. After the training process completed, the researcher would set up the UI (Figure 5.2), and allow participants to view the classifier's output in real-time using live data from the participant's Muse device. Participants would then see the probe's *accept* or *reject* classifications using the live data from their headset.

Free-form exploration ensued, as participants experimented with the device. During the experimentation phase, if subjects had not attempted to cause the device to fail already, the interviewer would ask, *"How can you be sure this device really works?"* This question was meant to encourage subjects to test the device's limitations, and to think aloud as they discussed why particular tests might be informative. The interviewer would then ask subjects to reflect on the device's limitations, and why such limitations might exist. *"Are these limitations intrinsic to the physical device? To brain-based authentication?"*

Finally, the interviewer would use the prior conversation to segue into a larger discussion

about brain-based authentication. *"Do you think the brain can be used to log people in?"* After some discussion, the interviewer would ask *"In your own words, what is authentication?"* This question would occasionally cause subjects to re-evaluate their prior answer. Finally, the interviewer would ask subjects what they believed the "ultimate" form of authentication might be. Through follow-up questions, this discussion would typically lead to the philosophical questions of personhood. *"What makes* you *you?"*

After some discussion, the formal interview would conclude, and the participants would remove the Muse device from their head.

5.3 Results

Participants experiences in using our technology probe as an authenticator were varied. Four particpants fuond the authenticator worked well for them, consistently accepting their readings when they were wearing the headset and consistently rejected readings when the headset was taken off, or when the experimenter was wearing it. On the other hand, four particpants found the authenticator did not work reliably for them. In two cases, the authenticator constantly rejected particpants regardless of whether it was on or off. In the other two cases, the authenticator consistantly gave mixed or ambiguous feedback when particpants wore it.

Regardless of the probe's accuracy, its ability to function at did lead subjects to discuss how the system could work, and how it would map to notions of identity. The remainder of this section outlines how such conversations emerged, and reviews participants' major beliefs about the ground truth for identity.

Failure modes.

Discussions about the nature of the self often started with discussions about how the authentication scheme might fail. In some cases, these limitations were perceived as cheiefly relating to the system's implementation. For example, one subject mentioned an opacity intrinsic to the black box algorithm used to perform the authentication.

> We don't know it's not working until it's not working. If something is inaccurate, sometimes 20% and sometimes 5%, there is uncertainty about what you dont know, not knowing what you dont know. *(Elizabeth)*

Here, Elizabeth refers to the difficulty of knowing when a false acceptance has occurred in such a system, and difficulty correcting for these mistakes. She draws this difficulty to a fundamental lack of information about possible inputs into the algorithm. We return to these notions of opacity, and their possible consequences, in the discussion. Other subjects brought up other possible shortcomings of the authentication scheme stemming from the changing nature of the brain.

> In cases of severe trauma or damage, like someone who drowns and is deprived of oxygen for some amount of time, or if you go through some major life trauma that's really awful, that can potentially change the brain. *(John)*

When asked whether it would be appropriate for a brain-based authenticator to reject a participant in such a situation, John reported that it would be. "Maybe that's colored by my - I work with computers, and I think that's a reasonable way for a system like this to work." Alex raised a similar point about brain changes affecting authentication, and also indiciated that it might be reasonable for the system to reject a user in such a case.

> It's a constantly changing fingerprint, so we have to continually update that... If the changes are so drastic, should you be allowed to access your bank account? When you take that acid trip, and - I had this big, life-changing trip and I can't log into shit anymore. "I'm a better person, I tell you!" *(Alex)*

In both of these examples, the changing nature of the brain brings difficulties for using the brain as a stable identifier. However, John did not indicate the changing brain would indicate a changing *person*. Instead, the system would reject the new person because of its reliance on the brain, and in spite of the fact that their personhood did not change. In contrast, Alex posits a scenario in which a transformative life event does change some aspect of his identity, making him a "better person." We return to these complementary meanings of a changing brain in the discussion.

Relationship between passthoughts and identity

To nudge subjects toward discussions about selfhood and its relation to biometric identifiers, the experiment would typically use discussions about the device's shortcomings or limitations to ask how a better system might work that overcomes these limitations. In so doing, the experimenter would ask subjects to draw comparisons between passthoughts and other authentication methods, both traditional (e.g. passwords or car keys) and inherence-based (e.g. fingerprints).

1. Skepticism

   All subjects expressed a high degree of skepticism that the passthoughts system we presented them could reveal identity in any true way. However, the reasons they gave for this skepticism were various.

   Following Dumit's work on the neuroessentialist belief that the nature of identity is fundamentally in the mind [**Dumit2004**], we expected subjects to believe passthoughts captured some truer token of identity as compared to other biometric or authentication paradigms. To our surprise, only one subject, Joanna, believed passthoughts captured something close to selfhood (discussed further under "Self as brain," below). Interestingly, this subject had an academic background in neuroscience prior to her career as

an engineer. However, informed by her training, she felt that current scanning machines had signals too noisy to reliable detect neural correlates of identity "anytime soon."

Other subjects gave more existential reasons for doubting that passthoughts meaningfully captured selfhood. For example, Alex felt that the EEG-derived authentication tokens were "basically a string" (e.g., a string of chracters, a common datatype in programming languages), and in this way similar to passwords. Alex's implication here is that the resulting string is crackable, and effectively a knowledge factor (to authenticate, one must only know a string that makes the authenticator output an *accept* label). Alex felt brain-based authentication schemes has little to do with identity for this reason.

Joe also sidestepped questions of identity in referring to the way passthoughts works. In contrast to Alex, however, Joe found the brain-based signals to be simple "learned responses," not tokens that meaningfully relate to a person's identity.

> I think it's just a learned response kind of thing, completely orthogonal to general questions of identity. You know, if I take 1000 pictures of you, track what you smell like, I guess that's latching on to some aspect of a fundamental "you," but I'm not sure that reflects who you are... *(Joe)*

Joe's analogy to photos and scents gesture toward the piecemeal nature of current biometric identifiers. Aspects of body are taken as evidence for personhood, which may "latch onto some aspect of" personhood, but does not reflect it in an essential way. (Joe later expressed that "mining" a true self is impossible, discussed further below).

2. Biological identity beyond the brain

While some participants did believe our brain-based authentication mechanism related to selfhood, these participants did not generally believe our brain-based scheme was particularly unique. Terrance, for example, compared our scheme favorably to that of "trust in some social institution." Although he did not specify which particular institutions he had in mind, he mentioned social security numbers, birth names and credit card numbers as other examples of authentication tokens. Though Terrance believed brain-based tokens were "more fundamental" than these identifiers, he did not believe they were uniquely well-suited to revealing selfhood, except that they were difficult for people to control directly.

> I can't even influence my brainstates, so how's someone else supposed to do it? But again, any organ like that would be equally good. *(Terrance)*

On one hand, notions of control gesture toward the possibility for coercion (i.e. forcing someone to reveal their authentication token), a point we return to in the discussion.

> On the other, the notion of the brain's non-specialness (i.e., that other sources of data could also be used to establish personhood, so long as they met particular criteria) mirrors in other subjects relational/longitudinal views of identity (below). We return to the consequences of such beliefs in "Implications for Design."

The "ground truth" of identity

Discussions about the relationship of passthoughts' implementation to the ground truth of identity caused many subjects to reflect on their own beliefs about identity. Mary explicitly did *not* feel the brain was the source of identity. In fact, she called out and critiqued popular narratives around brain's uniqueness:

> There is some special social narrative put on the brain, right. That the brain is, to use this phrase, the seat of your soul, and all that stuff. So I think that at scale, most people would ascribe some sort of significance to the brain, as a vector for this authentication mechanism generation - whereas I might not. But that's also just because I know - I think - that the brain is not any more special than like any other part. *(Mary)*

So much for uncritical engineers! Some subjects, such as Joe, refused to commit to particular notions of selfhood, stating such notions were impossible to "mine."

> We will never be able to mine the base of the self... I could record all my thoughts, my experiences, and plop someone into it, but there's no way to experimentally verify that it is my experience.

Joe raises a difficult epistemological problem in verifying that a given experience is truly what some other person experiences (probably this has precedent in philosph. should look for refs). However, implicit here is the belief that experiences and thoughts do equate with selfhood, gesturing toward a notion of selfhood expressed within the individual. On the other hand, some subjects posited a more social, relational view of selfhood. Mary referred instead posited a relational, social notion of "true" idenity, one with no particular relation to the body.

> I know this is not convenient for your brain thing, but I feel it's more actions... It's interesting, I went home a few months ago... When I go to Kanasas, no one thinks I'm charming. Everyone things I'm this realy type A, super intense individual. But here [in San Francisco], people say I'm charming. *(Mary)*

Mary's continued to explain that the true nature of self depended on where you were, and with whom you were interacting, drawing on her father's tendancy toward shyness around people he did not know well. John proposed a similar explanation.

> Your self is almost informed by other people, and the way they see you, and the way you see them, and the way you experience things throughout your life. I don't think it's spirtual thing, I think it's more like you've been writing in this journal for a really long time, and if you read it, all the way to the very end, the self is like the takeaway at the very end. *(John)*

Both John and Mary refer to the longitudinal and dynamic nature of identity, and implictly contested notions that a true or essential self might be derived from the body. The experimenter asked both subjects whether a system that detects their notion of identity could exist. Both expressed that it could, and in some sense, already did.

> The actions you take, the jobs you go to, the people you are interacting with, that is now, for the first time in history, written down in minutae. You can say, "this person's fingerprints are completely worn off," but you could predict if it's the same person based on their future actions. *(Mary)*

Drawing on the language of ubiquitous data collection, Mary referred to her own experiences as an engineer in making her aware of how pervasive such data collection is. Her vision of selfhood as captured in these "minutae" speaks to many existing theories of internet privacy like what??? why bring theories up at all, incl goffman - if these folks don't know about those theories, why mention them at all, esp in the results (vs discussion)?.

Dissenting views

1. Self as brain

   Two participants gave different views on the ground truth for identity. One participant, Joanna, offered a brain-based and biological explanation for the mind. Joanna felt that the brain was a truer source of identity than other parts of the body, though she noted that current devices, consumer and research-grade, were too crude to capture such an identity meaningfully.

   Joanna believed that the brain was the essential nature of the self. Though the brain "encountered input" from the body and world, it was the brain's "processing" of this input that gave rise to a self. In her explanation, Joanna drew on both her academic training as a neuroscientist, and also on her personal experiences with friends and family suffering from mental illness. She believed that changes to the brain during mental illnesses caused fundamental changes, such that these people were different selves.

   Interestingly, this subject was formerly trained as a neuroscientist, a group whose beliefs about the brain and self have been well-studied in the past [**Dumit2004**, **Vries2007**]. The fact that Joanna's beliefs were quite different from the other software engineers in our study raises intereting questions about how other technical practioners

might encounter notions of selfhood, or the brain, and how these beliefs may manifest in a future of more pervasive brain-computer interfaces. We return to this point in the discussion.

2. No true self

   Another participant, Terrance did not believe the self existed in any real or true way. He believed that "there is no fundamental difference between the self and the rest of the world" and that selves are defined "by convention" in social circumstances. Specifically, he believed engineers such as himself actively create such conventions through the artifacts they build.

   > Even if there's no logical bottom [to identity], we will create one.... we will find a way *(Terrance).*

   Terrance's response here gestures toward a reflexivity of his own role as an engineer, a view in which entities in the world are actively shaped through technical practices. At the same time, Terrance's response indicates a level of determinism, or resignation, that notions of selves will continue to feed themselves forward through technical artifacts. We return to this point in the discussion.

5.4 Discussion

This study focused on the beliefs of software engineers. We found that these engineers were well-positioned to be critical about the technical artifact we showed them. The system's highly technical interface and only semi-reliable performance encouraged them to enter their professional mindset as engineers, as they actively prodded at both the device and the ideas it embedded.

In their critical? interactions with our probe, participants generally acknowledged their unique subject position as engineers, either in their unflattering perspective on large-scale data collection practices, in their immunity to popular narratives around the brain, or in their reflexivity about their ability to construe concepts in the world through the creation of technical artifacts (e.g. terrance thought self didnt exist, would keep getting made as engineers "found a way").

To our surprise, we did not encounter the "neuroessentialst" view of self-as-brain found in prior work on neuroscientists [**Dumit2004**, **Vries2007**] except with one subject who had academic training in neuroscience. Prior work focusing on academic neuroscientists may not translate to technical practioners generally. As BCIs become more commercially accessible, and thus more amenable to experimentation by non-neuroscientists, future work should critically re-evaluate past work on academic BCI, with a particular sensitivity toward the consequences that beliefs about selfhood might have for technical practice. Since most emerging BCI technologies emerge from academic neuroscience [**Dumit2004**, **Vries2007**],

designers and researchers should be careful not to assume that past studies on the beliefs of neuroscientists will generalize to other technical communities.

The remainder of this discussion aims to trace through the main conceptions of self/identity we uncovered during our study. In the following section, we use these notions of selves to motivate future designs for authentication systems.

Various selfhoods

We present here notions of identity that are potentially overlapping, such that no catagory implies a whole or totalizing belief on the part of our participants. Rather, we acknowledge that each of these beliefs left room for one another. We often encountered a coexistence of these beliefs among our participants.

1. Dynamic selves

   In evaluating passthoughts, participants typically grappled with the reality of a brain that changes over the course of one's life. Through their reflection on this topic, participants expressed a belief that people themselves change over time, contrasting with traditional inherence-based tokens, which are explicitly designed to be static throughout a person's life [**Bonneau2012**].

2. Relational selves

   Four of our subjects expressed a belief that selves change depending on context, particularly where a person is and with whom that person is interacting. (Though subjects did not discuss it as such, these beliefs relate somewhat somewhat to Irving Goffman's dramataleurgical analysis of social behavior as performance [**Goffmann1959**]). slice the goffman ref, or unpack and make relevant to design In any case, these beliefs contrast with inherence-based authentication tokens, which are by design insensitive to contextual conditions.

3. Data selves

   In describing something that maps closely to a ground truth of selfhood, three subjects referred to the aggregate of all digitally captured data over time. These participants often made analogies to written records in explaining this connection (John explicitly mentioned a "journal," where the self is "the takeaway at the end"). This notion of a selfhood-through-data relate to some past work in quantified self [**Nafus2016**], though this work focuses more on self-tracking practices than authentication in particular. We return to this connection to past work in the following section.

4. No selves

   Two of our participants believed the self was an untenable construct, either because they believed it was not real (Terrance) or impossible to verify experimentally (Joe). However, Terrance believed that notions of self could (and will) be constructed by

engineers. In the following section, we discuss further what it might be like to build an authentication mechanism that does not make commitments about sefhood.

5.5 Implications for design

In our discussion, we surfaced tensions between what inherence authentication does (i.e., verify personhood), and what authentication systems would have to do in order to classify *personhood as engineers understand it.* In reflecting on these tensions, we present here some possible designs more ammenable to the notions of selfhood we surfaced in our study. We quickly explain each method, and describe how such a system may motivate future research.

Sort of interesting to me

Here's some stuff that's sort of interesting to me, some work already speaks to it. I feel I should talk about in the interest of mapping sense of self to alternative authentication technologies.

1. Social authentication

   motivate with discussion

   - participants felt self had to do with how you interact around others

   quick explanation from sec lit

   - socially contextual authentication, who you are around and who knows you [**Das2017**]

   why/how interesting for design/future work

   - could help map to **diff** notions of selfhood and autonomy, e.g. in cultures where family or community autonomy is emphasized over individual autonomy [**ur2013cross**].

2. Continuous, multimodal authentication

   motivate with discussion

   - views of self as dynamic, captured in longitudinal data
   - relational, action-oriented, based on *what you do* rather than *what you are.*
   - mary gestured toward existing systems of massive data collection,
   - Terrance gestures toward "control," difficult to control is good (like brain)

   quick explanation from sec lit

   - "what you do" .. all sorts of data, all digital traces
   - en masse used to distinguish you from others

why/how interesting design/future work

- plays with privacy, tension between privacy and security (autonomy), similar to what we play with here using our "mind-reading" probe
- Algorithmic opacity/transparency came up
  - significance to continuous auth, which is almost always probabilistic [TODO comb thru those prob/continous auth pdfs]

Real interesting to me

Here's a concept I'm really interested in, no work out there on it that I know of.

1. Authentication without the self

   motivate with discussion

   - Terrance didn't believe self was real, "no logical bottom"
   - Joe thought self could not be verified

   how/why interesting for design/future work

   - Think hard about whether auth systems should even care about the self
     - Is (bodily) selfhood even something we want or need?
     - Could reference Proudhon's "invention of the self," a technology that makes markets work
   - What would it look like to provide authentication without a sense of personhood?
     - Passwords are like that, just about access!
     - Other things that are similar? Possession factors e.g.?
     - Group-based authentication, questions primacy of individual?
     - Systems with no static identities, only changing psuedonyms [must be prior work on this]?
   - What upsides, why apealing beyond theory?
     - For research?
     - For probing beliefs?
     - For questioning dominant assumptions about what computer security should acheive / accomplish?

Selfhood beyond authentication

For what other applications is personhood is relevant? Personhood in software beyond authentication?

- captchas that are behavior based?...
- profiling systems, e.g. recidivism?...
- Implications of this work for those systems?
  - TODO what designers can be sensitive to/how..
  - TODO blurring of the line btwn systems that track behavior + systems that track identity....?.....hmmmmmmmmmmm.....kind of interesting no??

5.6 Future work

While our probe succeeded in surfacing beliefs about personhood, our study was limited by the short duration of our probe. Our sensing device did not have the battery life to record for several hours, and we experienced intermittent connectivity issues when walking outdoors. A device now available [**Optical2017**] would allow us to deploy a future probe over hours or days, enabling a study in which participants could build a meaningful relationship with the device over time and in the context of everyday life [**Gaver1999**].

One interesting direction for future work could be comparative between engineers and non-engineers, looking at the origins of any differences that arise. more...?

longitudinal data selves, what about china's social credit system? private-sector surveillance on the net as selfhood, alternative to state-given identity? HMMMMM now thats an interesting one.

5.7 Conclusion

- Authentication and property / Authentication in society?
  - Authentication's primary role in controls access to property?
  - biological identity / consequences of these beliefs, wrt property, access to services, criminality. talking to the ppl who build these consequential systems. a motivation for a whole new paper

Chapter 6

Conclusion

surveillance, simone browne, about power

so i went through nitpicking all these theories, but..... what can we add to conversation that will substantially help to address these critiques across much research?

High level takeaways from this discussion section. Answer the paper's question: what are the limits?.

Brainscanning, autonomy, control

TODO A future for privacy and cybersecurity

now with VR even more intimate bci promises yet another intimacy, look at those side-channel attacks and so on, done with P300 and now there's a startup with a P300 api [], once the stuff of fiction []

people will continue to build increasingly hi-res models of human bodies in space;, and human environments I argued that these models will in general will be informative wrt *the mind*, producing what i dub *mind-computer-interfaces* (MCIs).

what 2 do now?

- It is now time to see if engineers believe mind is *senseable*
  - See how our theory matches up (or doesn't) with their beliefs
- One good starting place is the brain
  - But other wearable senosrs can also work .. heart is a good one, lots of connotations there, and may be diff btwn cultures!

WHY SO IMPORTANT?
TODO why this section