The Importance of Visibility for Folk Theories of Sensor Data

Emilee Rader and Janine Slaker, *Michigan State University*

https://www.usenix.org/conference/soups2017/technical-sessions/presentation/rader

This paper is included in the Proceedings of the
Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017).

July 12–14, 2017 • Santa Clara, CA, USA

ISBN 978-1-931971-39-3

The Importance of Visibility for Folk Theories of Sensor Data

Emilee Rader
Department of Media and Information
Michigan State University
emilee@msu.edu

Janine Slaker
Department of Media and Information
Michigan State University
slakerja@msu.edu

ABSTRACT

Sensor-enabled wearable devices and smartphones collect data about users' movements, location, and private spaces and activities. As with many ubiquitous computing technologies, this data collection happens in the background and appears "seamless" or invisible to the user. Despite this, users are still expected to make informed choices regarding consent to data collection. Folk theories are sets of beliefs and understandings that arise informally and guide decision-making. To investigate folk theories regarding sensor data collection that might guide privacy self-management decisions, we conducted qualitative free list activities with 30 activity tracker users in which we asked them to list "information that an activity tracker knows". We found that folk theories regarding the data that activity trackers collect depend on interactions between the users and their trackers that provide visibility into dependencies among data types, evidence about what trackers are able to record, and feedback that inspires speculation about how trackers work. Our findings suggest opportunities for designing interfaces that intentionally support the development of folk theories about how sensor data are produced and how they might be used, which may enable users to make more informed privacy self-management decisions.

1. INTRODUCTION

Ubiquitous computing systems that incorporate a wide variety of sensor technologies are an increasingly common part of everyday life for many people. In particular, wearable devices like smart watches and fitness bands, and smartphones carried in a pocket or purse, have been widely adopted. All of these devices include embedded sensors that engage in continuous data collection, and are capable of producing inferences that users consider "extremely private" [39]. For example, in February 2016 a Reddit user posted heart rate data from his wife's Fitbit activity tracker to enlist the community's help in troubleshooting what he believed was a malfunctioning device. Instead, he found out from other users that what he had noticed could actually be valid data indicating that his wife might be pregnant (in fact, she was) [17].

User awareness and concern regarding data sharing and use often receive more attention in the security and privacy literature than data collection does. Privacy concern has been shown to depend on contextual aspects of sharing and use [30, 33], and encouraging people to think about different possible audiences and uses can affect how concerned they are [23]. However, data collection practices have also long been recognized as related to privacy. This was acknowledged in the original 1973 report on which the Fair Information Practice Principles (FIPPs) were based [52][1]. It was again recently emphasized in the 2012 "Consumer Privacy Bill of Rights" report issued by the Obama administration [47], which includes the directive, "Consumers have a right to exercise control over what personal data companies collect from them and how they use it."

People are expected to be able to self-manage their privacy by making decisions about what systems to use and what kinds of data collection to consent to [45]. This approach assumes that all users are able to think and behave correctly, and in an informed and rational fashion, which is not realistic [22]. An approach adopted by security and privacy researchers regarding how to understand users' choices and behavior focuses on folk theories related to technology use [2, 53, 57]. Folk theories are beliefs, analogies and explanations that guide people's behavior, which develop and evolve through everyday experiences. Folk theories about how technologies work form even when details about the inner workings of the technologies are invisible to users [11, 37]. By investigating folk theories related to sensor data collection, we can gain insights into how everyday interactions with sensor-enabled systems support their formation. We can also find out more about what guides users' privacy self-management decisions and behavior regarding these systems.

We conducted a qualitative study focusing on folk theories about data collected by activity trackers, defined as smartphone apps and standalone devices that support fitness-related data collection (movement, heart rate, etc.). These devices and smartphone apps are an example of sensor-enabled technologies that have achieved wide, mainstream use. They also have an interface that provides information to users about the data they collect; seeing step counts and other health and fitness activity information is part of the reason why people use them. Our focus is considerably more narrow than studies like Wash's folk models of threats [53] and Yao et al.'s folk models of online behavioral advertising [57], and more like Kempton's study of thermostats which focused on a single application [22].

We found that participants' folk theories conceptualized types of data their trackers were collecting as if they were either manually *entered* by the user, directly *measured* by the tracker, or *calculated*

[1] "There must be a way for an individual to find out what information about him is in a record and how it is used." [52]

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

from other data the tracker had collected. Participants' folk theories had developed through interactions with their trackers that provided visibility into dependencies between different kinds of data, such as values presented in the interface that increase at the same time, visible heart rate sensors, and step counts shown by the tracker which did not seem to match participants' perceptions of their activity. However, participants' folk theories had not incorporated the idea that their data were estimates or inferences. This precluded speculation and reasoning about how raw sensor data might be useful for other purposes outside the context of activity tracking that could reveal information about activities and personal characteristics users might not want to disclose. In other words, these folk theories would not support users being able to consent to uses beyond the context of activity tracking.

With this paper, we make several contributions. We present findings about folk theories related to sensor data collection in ubiquitous computing systems that provide insight into what users are aware of and can reason about, in their own words and from their own perspective. We highlight the importance of making aspects of the data visible in the interface, and of designs that encourage users to speculate about the origins of their data. And, we present design implications for ways activity tracker interfaces might better support speculation, and thereby the formation of folk theories that help users reason and make decisions about privacy.

2. RELATED WORK

2.1 Folk Theories

Folk theories are "ways of understanding" that help people interpret phenomena they encounter in the world [12]. They are based on experience and interaction rather than formal instruction, and are often shared across people [21, 22]. Folk theories for specific technologies arise out of users' everyday experiences with those technologies [4, 22]. This means they sometimes vary from person to person, and are often incomplete and inaccurate from an expert point of view [11, 4]. Folk theories help users generate explanations [18], guide inferences they make about cause and effect [4], help them reason about what technologies are capable of [35], and influence their choices and decision-making with respect to those technologies [53]. Folk theories are also sometimes called folk models, or mental models, and are elicited through qualitative investigations that involve interviews and hypothetical scenarios [53, 56], activities like sketching [19, 57], and prompts such as photos [35] or specific app permissions [24] that participants react to. Many studies have found that folk theories held by experts are different from those held by non-experts; for examples, see Asgharpour et al. [2], Kang et al. [19], and Yao et al. [57].

In a widely-cited study, Wash [53] investigated folk theories that non-expert computer users have for security threats like hackers and viruses, and argued that they have implications for whether people believe particular pieces of security advice are important to follow or not. Kang et al. [19] investigated folk theories of the internet, and found that non-experts drew simpler diagrams of locations where data is stored online and where it travels to, whereas experts drew more complex diagrams with more parts and components. Both experts and non-experts knew that their data is shared with companies that provide services to them. But beyond that, all participants expressed a lot of uncertainty. A more recent study about online behavioral advertising found that folk theories involve beliefs about agency, including which entities are involved in tracking users, and where the data are stored. These ranged from a browser-pull model where the browser is responsible for storing all user data and obtaining relevant ads, to a more technically accurate model that involved both first- and third-parties [57]. Folk theories about RFID, which is sensor-related technology, have been found to be partially correct as well: the most common folk theory involved the idea that RFID tags are small devices that can hold a little bit of information, similar to a magnetic strip or a barcode [35]. And, related to activity trackers in particular, Yang et al. [56] found that many activity tracker users had engaged in "ad-hoc assessment" of how their trackers worked while they used them, which resulted in feelings of frustration related to perceived inaccuracies in their activity tracker data. Yang et al. recommended that users should be provided guidelines for how to determine the accuracy of their devices, calibration mechanisms so trackers can better adjust to individual variation, and training resources that explain what trackers measure in order to increase understanding. These studies are all examples of investigations into folk theories of technology use, and serve as background about the approach we take in our study, and the kinds of insights about users, and about design, that this approach supports.

2.2 Wearables, Smartphones and Privacy

A classic early paper about wearable sensor technology and privacy describes the Active Badge system [51]. This system was developed in the early 1990s as a proof of concept location tracking system for a research organization. It used RFID technology to track users' locations via wearable badges that could be detected by a sensor array in the building. Many people found the system to be useful, in that it enabled them to find colleagues more easily when they wanted to talk to them. However, people also expressed concerns about privacy, mostly surrounding ways that the location data provided by wearing the badge could potentially be used, and by whom. This theme about data use comes up in many studies about sensor data and privacy. Even when users report that they feel the collection of the sensor data itself does not concern them, when asked to consider possible uses they are able to imagine harms that might result if the data were used improperly.

For example, Klasnja et al. [23] asked participants about privacy concerns related to their use of the UbiFit wearable fitness prototype over a 3-month period. None of the participants were concerned about the idea of data collection, because they didn't think the data were sensitive on their own. However, they had concerns about the use of location data, and raw audio data. Similarly, Raij et al. [38] showed participants who used the AutoSense system for three days visualizations of the data that had been collected about them. Participants reported increasing concern about more sensitive kinds of data (e.g., less concerned about physical activity, more concerned about conversations and stress level). In contrast, Rapp et al. [39] and Motti and Caine [29] both found that users of commercial activity trackers did consider the data that was collected to be extremely private. But, in both studies participants remarked about concern due to a feeling that they were not necessarily in control of how their data might be shared with third parties and subsqeuently used.

Other researchers have focused on the issue of smartphone apps accessing and potentially sharing user data with third parties, or data "leakage". Shklovski et al. [44] interviewed smartphone users about this, and found that it is a source of concern for users. Balebako et al. [3] took this a step further by notifying participants when smartphone apps were accessing data; participants were surprised about how often this happened, and more concerned about it than they were before they were made aware of it. Almuhimedi et al. [1] also focused on smartphone notifications in their research, and used the notifications to provide a means of awareness and control over smartphone app data access and use. In their study, the notifications showed users information about how many apps had accessed different types of smartphone data in a specified time period. This style of intervention assumes that drawing users' attention to data sharing and use by informing users about how many times different apps were accessing their information would raise concern and "nudge" users to take action. Just over half of their participants in this study made changes to app permissions as a result of the intervention.

Shih et al. [42] also studied smartphone users' privacy concerns and willingness to share data, via a custom app they created. The app was designed to measure participants' privacy preferences regarding app usage of personal data by asking them questions periodically over the 4-week period of the study involving different combinations of app types, data types, and usage purposes. Users were least willing to share information when more details were given to them like the name of the app that was using the information, and what the app was going to do with the information. In other words, providing more detail about use was associated with less willingness to share the information.

We focus in this paper on an application of sensor data collection that is commercially available and in mainstream use: wearable activity trackers and smartphones used for the purpose of activity tracking. This sampling frame enabled us to recruit participants who had already been using sensor-enabled devices and smartphones for their own reasons, some of them for a number of years. Activity trackers already support some form of user interaction with the data they collect, which presents researchers with an opportunity to study folk theories of sensor data collection that have developed over time in actual use, rather than as a short-term research intervention.

Unlike previous work on folk theories of technology related to security and privacy, with the notable exception of the study by Poole et al. about RFID [35], our study specifically involves sensor technologies. It is also different from work focusing on privacy awareness and concern related to data sharing and use, because it focuses primarily on data collection. And, it is different from many security and privacy studies in that we do not assume there is an objectively "correct" behavior that users must be measured against. Rather, we focus on understanding non-expert users' existing folk theories from their perspective, so that we can better understand what guides their behavior, and make recommendations for design to support the formation of more privacy-relevant folk theories.

3. METHOD

3.1 Data Collection

We conducted semi-structured interviews that began with a free list activity designed to elicit folk theories about what types of information activity trackers are able to collect, and about how they are able to collect that information. Interviews lasted roughly 60 minutes and took place primarily over the phone, with a few in person, during December 2015 through February 2016. There are several advantages to conducting phone interviews versus in-person interviews. Phone interviews allow access to participants in diverse geographical locations, maintain interviewee anonymity, and can decrease social pressure and increase rapport. Research on the two methods has not found either to produce data of compromised quality [31, 46].

The free list activity lasted about 12-15 minutes, and in every case took place at the beginning of the interview, right after obtaining consent. The remainder of each interview after the free list activity focused on participant thoughts and reactions regarding a series of hypothetical scenarios in which activity tracker data might used to infer other kinds of information about the user. Each participant received a $25 Amazon.com gift card as a thank-you for participating. This study was approved by our institution's IRB. In this paper we focus on just analysis and findings from the free list part of each interview.

Free listing is a method used by anthropologists to elicit concepts that are part of a semantic domain for a group of people. Free list activities begin by the interviewer prompting the participant to "list all the kinds of X [the domain] you can think of" [6]. The interviewer then follows up with additional prompts to clarify things the participant has said and elicit additional concepts until the participant runs out of concepts to list. The goal of free listing is to gather data about the structure of a semantic domain and the relationships between concepts within the domain, as understood by the participants [55, 48]. In other words, the intent is to understand the semantic domain from the participants' perspective, not to impart any external structure onto what participants have said. Items or concepts that are mentioned infrequently, or not at all, are not considered to be part of the semantic domain according to participants [36].

Free list activities are unlike other semi-structured interview techniques in that they elicit information about things which at least "in principle have a right answer which is universally true". Participants in a free list activity should feel like they are discussing facts about the world, "perceptions, not preferences" [7]. This is an important distinction for our study, because folk theories arise out of everyday experiences in the world [22]. Therefore, we used a method to elicit participants' knowledge and understanding of the world within the semantic domain of interest to our investigation, not their attitudes, opinions, or concerns.

The wording of the domain-specific prompt we used for our free list activity was *information that an activity tracker knows*. This prompt was specifically designed to elicit concepts related to the data activity trackers collect, without priming participants to use "data"-centric terminology or focus their attention other aspects of activity trackers and data collection introduced by the researchers. The prompt did not ask participants to speculate about what might be possible for activity trackers to infer about users, instruct participants to imagine things an activity tracker *might* know, or list information that other people (instead of a device or system) might be able to infer based on activity tracker data. We avoided prompts that might encourage participants to speculate, because this could prime them to think about something they had not considered before. We wanted to elicit their existing folk theories rather than encourage them to develop new ones.

Free list activities often produce information that is incomplete or ambiguous, because recalling all associations is difficult for participants to do [8]. Most of our participants began by listing concepts related to their knowledge of activity trackers in general, and as the activity progressed they made more specific references to the tracker that they personally used. We did not direct them to focus on specific features or technical capabilities of their own particular activity tracker; rather, the prompt was intentionally general to allow participants to describe using their own language what they understood about the information that activity trackers collect. After each participant finished making his or her initial list, the interviewer read the list aloud which helped the participant to generate items they had initially forgotten to include [40]. Additional follow-up prompts were used to clarify what the participant meant

ID	Age	Gender	Activity Tracker
P01	44	F	Fitbit Flex
P02	27	F	Polar Beat App w/heart rate band
P03	32	F	Fitbit Flex
P04	48	F	Fitbit (wristband)
P05	34	F	Fitbit (unspecified)
P06	39	F	iOS Health, Move Apps
P07	30	F	Pacer App
P08	42	F	Virgin HealthMiles Pedometer
P09	32	F	Fitbit Charge HR
P10	38	F	Fitbit Charge HR
P11	23	F	Fitbit Charge HR
P12	39	F	Fitbit One
P13	40	F	Samsung S Health App
P14	24	M	Fitbit Flex
P15	36	F	LG Health App
P16	29	F	Fitbit Charge HR
P17	24	F	Google Fit App
P18	25	F	iOS Health, MyFitnessPal, WeChat Apps
P19	25	M	Argus App
P20	35	F	Fitbit Charge HR
P21	40	F	Fitbit Charge HR
P22	32	F	Samsung S Health App
P23*	34	F	Fitbit Charge HR
P24*	24	F	Fitbit Flex
P25	34	F	Fitbit (unspecified)
P26*	28	M	Fitbit One, heart rate band
P27*	33	M	iOS Health App
P28	24	F	NexTrack App
P29	36	M	iOS Health App
P30	25	M	Fitbit (unspecified)

Table 1: Participant characteristics. ID numbers with an asterisk (*) indicate participants who were no longer using an activity tracker at the time of the interview.

by the terms they listed (e.g., "What do you mean by X?" where X was the term mentioned by the participant). After the free list activity was complete, the interviewer asked the participant additional follow-up questions about the items they had listed, to elicit associations between different terms participants mentioned, and between the terms and tracker-related activities and use. For example, a follow-up question frequently asked was, "Can you tell me how you think it knows X?" (e.g., can you tell me how you think it knows steps?). The "how" prompts allowed us to elicit participants' understanding about dependencies and causal relationships between different types of information collected by their activity trackers. In the follow-up prompts, the interviewer took care to refer to concepts introduced by participants using the same terminology that they had used.

3.2 Participants

We recruited participants who were current or former users of activity trackers, which we described in our recruiting materials as *wearable activity trackers and mobile sensors that automatically count steps, like Fitbit or the Moves app*. We included both wearable devices and smartphone apps in our sampling frame because they are used for similar purposes (e.g., step counting) and collect similar data (e.g., via accelerometers). We advertised our study using snowball sampling on Facebook and email sent to a paid research pool organized by our institution. The paid research pool at that time consisted of about 3700 active users from the local community surrounding a large midwestern university. We combined these two methods of recruiting to obtain a more diverse sample in terms of geographic location [25] and demographic characteristics [41]. Roughly 60% percent of our sample came from snowballing. Friends and family members of the researchers were ineligible to participate, as were undergraduate students, and anyone who reported on the screening questionnaire that they had received training or worked as a computer programmer, software engineer, or in some other IT-related position. Folk theories of a variety of technologies have been shown in previous research to differ between experts and non-experts [2, 19, 26, 34]; the folk theories of experts are more complex and use more specialized vocabulary. We excluded technology experts from our sampling frame because we expected that they would be more familiar with how the underlying technologies work. Also, expert users may view privacy self-management differently than non-experts do. We also chose not to recruit from enthusiast venues like Quantified Self forums or to target early adopters, because we did not want to bias our sample towards self-tracking experts who might be more knowledgeable about how sensor data are produced.

Our sample consisted of 30 participants (80% female; mean age = 32.5; age range = 23–48) who lived in areas across the U.S. (e.g., Illinois, California, New York) in both urban and suburban settings. Many were administrative assistants, homemakers, and worked in research-related professions (lab managers, analysts). We also interviewed participants who worked in healthcare, state services, law and business development. Market research shows that women are more likely to own an activity tracker than men [16], and also more likely to volunteer to participate in research when online recruiting methods are used [14, 32]. While our sample was primarily female, we actively looked for evidence of differences between men and women in our data, and did not find any. All participants were current or former users of activity trackers. Eleven participants had been using a tracker for one or more years; six for 6-12 months, and four for 1-5 months. Nine participants did not report how long they had used an activity tracker. Twenty to thirty participants is a reasonable sample size for free list activities that involve a small or well-defined semantic domain, according to Weller and Romney [55]. Table 1 presents a summary of some of the characteristics of our participants and the trackers they used.

More than half of our participants used wearable devices created by Fitbit, which monitor activities ranging from step counts to sleep patterns, and provide additional information about users' activities such as active minutes and calories burned. The Fitbit Charge HR (the most popular among our sample) is distinct because it continuously monitors a user's active heart rate. Only two of the 19 participants who used a dedicated activity tracker device did not use a Fitbit. Eleven participants used an activity tracker app on their mobile phones, without a separate wearable device. These apps use sensors within the phones to track steps and other data. The Samsung S Health app uses a similar technique as the Fitbit Charge HR for measuring heart rate in which the user places her finger onto an optical sensor (located besides the phone's flash) and LED light is reflected onto the skin to determine the rate of expansion and contraction of the user's capillaries. We consider both dedicated activity tracker devices and smartphone apps to be "activity trackers" for the purpose of this study, because our participants self-identified them as activity trackers, and because according to our participants both perform similar functions and collect similar kinds of data.

3.3 Analysis

Interviews were digitally recorded and transcribed, and the transcripts were divided into two files for analysis: one containing just the initial free list activity and another for the remainder of the semi-structured interview. We analyzed the free list transcripts using an iterative, inductive coding approach which identifies themes

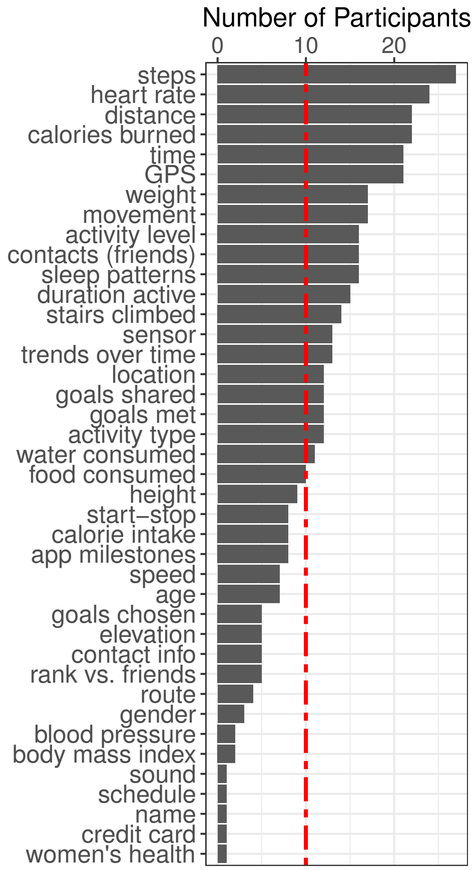

Figure 1: Histogram of the number of participants who mentioned each data type. The bars reaching beyond the dotted line are data types that at least one-third of participants mentioned.

to address "relationships of similarity" [27]. We began by standardizing phrases or statements made by participants about the same types of data so that we could generate counts of how many participants mentioned each data type [55], as is typical for analyzing free list data. While participants generally used similar words to refer to similar data types, there were some differences. For example, most participants talked about "movement", but we also coded terms like "vibration" (P16) and "jarring sensation" (P27) as movement. Participants mentioned 40 different data types; the mean number of data types they mentioned was 14.2 (*SD* = 3.86). The most commonly mentioned data type was steps taken (27 out of 30 participants); however, only 17 participants mentioned movement, which all activity trackers record to some extent. Figure 1 shows a histogram of all of the data types participants mentioned.

In addition to coding for data types, we coded for statements that participants made about *how* activity trackers come to know the information. We standardized the verbs participants used to talk about how the tracker knows, so that we could determine how many participants used these concepts and analyze which verbs were used in conjunction with the types of data that were mentioned. For example, both P14 and P22 talked about how their trackers know when they are engaged in a higher level of activity. The italicized sections of the quotes below indicate the connection each participant is making between the verb and data type they mentioned:

> If I'm moving frequently for 20 to 25, 30 minutes, I think *that gets tallied in the active minutes section*. (P14)

> I mostly run on the treadmill. So when I use the running in my app, it's not literally track[ing] the GPS so it seems like I'm not running at all because it tracks with the GPS. And so, this cannot be taken as moving, so *it's not counted as moving*. (P22)

The verbs in both of these examples, "tallying" and "counting," were standardized as "counting". Overall, participants used 32 different standardized verbs to describe how the tracker knows different types of information, and the most commonly used verb was "tracking" (15 participants), followed by "inputting" (11 participants).

After the final data collection and coding, we constructed two summary matrices [28], one consisting of transcript excerpts containing co-occurrences between different data types, and the other consisting of co-occurrences between data types and descriptions of data provenance. The matrices included only data types that were mentioned by at least 10 participants. We identified the data type(s) in each excerpt, any relationships between the data types (e.g., one information type being based on or calculated from another), and descriptions of data provenance. We used this rich dataset to generate visualizations of the connections and dependencies between data types, and identify higher-level patterns.

3.4 Limitations

The method and sampling frame we used have several limitations. We had a small sample that was selected for diversity, not generalizability. This means that our findings cannot be interpreted as statements about prevalence in a wider population of activity tracker users. Also, our qualitative data come from retrospective self reports. This is appropriate for the free list technique, but it means that we did not observe participants interacting with their activity trackers, or directly study the formation of folk theories as it happened. In addition, the data were collected by eliciting responses to a specific prompt we designed for the free list activity. There may be salient data types that participants did not mention due to the wording of the prompt and follow up questions. In particular, the choice to use a general prompt, and not to direct participants to speculate, means that we can't draw conclusions about folk theories for what activity trackers might be able to infer. Finally, because we did not ask about privacy concern as part of the free list activity, we can't use these data to connect the folk theories to specific concerns about privacy related to sensor data.

4. FINDINGS

4.1 Folk Theories about Types of Data

Our participants' folk theories about sensor data collected by activity trackers included three different categories of data types, differentiated by how they believed their trackers were able to collect or generate the data. These categories were not always technically correct compared with how activity tracker technology is actually able to generate the information provided to users, based on user documentation available from activity tracker companies and whitepapers about sensor technologies[2]. We first discuss relationships and dependencies participants described between the types of data they mentioned, and then use the pattern of dependencies to

[2]Fitbit, in particular, has extensive user documentation available on its website, help.fitbit.com, accessed on June 10, 2017.

Figure 2: The relationships between the data types mentioned by participants. Arrow thickness indicates the number of participants who mentioned a connection between a pair of data types. *Entered* data types are blue, *measured* are green, and *calculated* are red.

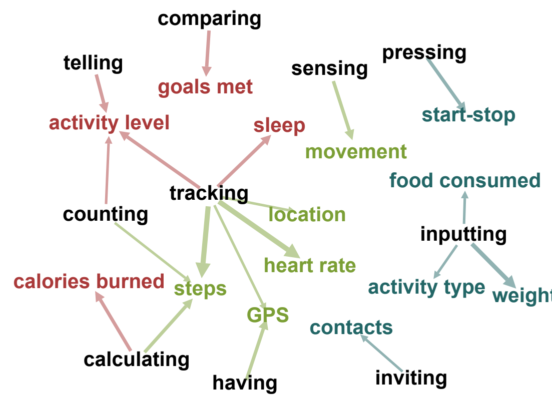

Figure 3: The relationships between data types and the verbs participants used to describe how the tracker records each data type. Arrow thickness indicates the number of participants who mentioned a connection between a pair of data types. *Entered* data types are blue, *measured* are green, and *calculated* are red.

illustrate the three higher-level categories of data that were present in participants' folk theories.

4.1.1 Dependencies Between Data Types

We identified dependencies between data types by focusing on statements participants made that indicated one data type was based on on another, for each data type mentioned by 10 or more participants. All data types participants mentioned are included in Figure 1; we focus here on dependencies between data types listed in that figure having bars beyond the dotted line, from "steps" to "food consumed". For example, P17 explained the relationships between the data types that allow her tracker to determine calories burned (emphasis added):

> I think, based on the metrics I've given it: My *age, height, weight*, so it knows all that and then it calculates *based on* my average *activity level*, how many *calories I've burned* for the day.

After identifying pairs of data types mentioned by each participant and the direction of the dependency between them, we created a network graph to visualize these relationships. This graph is presented in Figure 2. Arrows point from the *antecedent* data type (e.g., weight) to the *descendant* data type (e.g., calories burned). Only those pairwise relationships mentioned by at least three participants are included in the graph. Thicker arrows indicate that more participants talked about the existence of that relationship. Common relationships included movement to steps, GPS to location, and sensor to heart rate. Not all participants mentioned the same relationships between pairs of data types. For example, 9 participants said that steps were based on movement; however, 18 participants mentioned steps alone, without another data type.

When participants described how an activity tracker might know a certain type of information or described the relationship between a pair of data types, they often used verbs to describe the nature of the relationship. We created a second visualization (Figure 3) depicting co-occurrence between the data types and the verbs they mentioned. Arrows point from the verb to the data type, and the thickness of the arrows represents how many participants used a particular verb. For example, the verb "inputting" was used to describe how the tracker knows the user's weight. The verb "tracking" was used in conjunction with 6 different data types (steps, GPS, heart rate, location, sleep, activity level), indicating that participants used this verb in a nonspecific way to refer to something the activity tracker did to collect data. Some verbs were not used consistently across participants. "Counting" was used to talk about both steps (counting the total number of steps) and the user's activity level (what counts as an active minute). And, "calculating" was used to talk about both steps (calculating the number of steps for the day) and calories burned (calculating how many calories burned). Participants also talked about trackers "having" GPS, and "sensing" movement. Diagramming these verbs and dependencies allows us to analyze the cause and effect relationships participants described, and determine similarities between the data types participants discussed. This reveals characteristics of their folk theories.

4.1.2 Three Categories of Data

We identified three categories of data types our participants discussed: *entered, measured, and calculated* data. These categories emerged from the dependencies between data types mentioned by our participants, and the different verbs participants used to connect the data types. Table 2 shows all types of data that our participants identified during the free list activity, separated by category. Broadly speaking, *entered* data consists of data types that users manually input into their activity tracker interfaces. *Measured* data types are those participants described as being directly recorded, standalone phenomena. *Calculated* data types are considered to be generated based on other data types. These three categories are important for understanding participants' beliefs about the kinds of data collection their activity trackers engage in. We describe each of the three categories below.

Entered Data: Twenty-eight of our 30 participants reported that an activity tracker collects some information that users manually input, by entering it into a standalone tracker app or an app associated with their tracker device. All entered data types had zero antecedents (see the blue data types in Figure 2 that have no incoming arrows), and participants talked about them using verbs that indicated manual data entry. The entered data types include physical characteristics like age, gender, weight and height.

One entered data type that was mentioned by 8 participants, "start-stop", was not a characteristic of the user like weight or gender. Instead, it marked a transition between one activity state and another, such as being awake versus asleep. Participants talked about "pressing" a button (P12, P15, P24, P25), or "setting" (P05, P29), "turning on" (P05, P28) or "telling" (P01, P05) the tracker to enter sleep mode or to start or stop tracking an activity like a walk or a run. Only one participant mentioned entering information that was not directly related to fitness tracking, like name and credit card information. This is surprising, because activity tracker apps require users to create an account in order to use the service, and ask for information like name, contact information, and in some cases payment information as well.

Measured Data: All 30 of our participants believed that some data types, like heart rate, movement, steps and GPS, are direct measurements collected by the tracker. Measured data types are similar to each other in that participants mentioned virtually no antecedent data types in connection with them (incoming arrows in Figure 2; measured data types are green), or referred to them as being automatically detected by the tracker.

There are three measured data types in Figure 2, time, sensor, and GPS, that have no incoming arrows, indicating that participants believed these were not based on any other data types. However, there was some disagreement among participants about whether steps in particular had an antecedent data type. P16 provides a typical example of this:

> And yes, it definitely guesses by vibration as well, or by something like that, so it knows how many steps I take per day.

In this and similar examples (e.g., P19 said his tracker "record[s] leg movement"), participants were aware that steps are calculated based on movement. However, across our participants, steps were more similar to the other measured data types; almost twice as many participants talked about it that way. Seventeen participants did not mention an antecedent data type at all for steps, while the other 9 participants who mentioned steps said it was based on movement.

Calculated Data: Participants described calculated data types differently from entered and measured data types; all were described as having two or more antecedent data types. For example, in Figure 2, there are arrows from four different data types pointing to calories burned. P23 showed a more sophisticated understanding of this than most of our participants:

> So steps would then translate into miles traveled or some do specifically steps calculated and then they don't know, but they approximate calories burned, based on who knows what algorithm.

P23's perception that an algorithm is involved is actually fairly technically accurate. For example, the Fitbit help pages say that calories burned is "estimated based on the physical data you entered when you set up your account: gender, age, height, and weight," and "the activity recorded by your tracker"[3].

[3] `https://help.fitbit.com/articles/en_US/Help_article/1381` (accessed on July 10, 2017)

Measured		Entered		Calculated	
steps	27	weight	17	calories burned	22
heart rate	24	contacts	16	distance	22
GPS	21	activity type	12	activity level	16
time	21	goals shared	12	sleep patterns	16
movement	17	water consumed	11	duration active	15
sensor	13	food consumed	10	stairs climbed	14
location	12	height	9	trends	13
elevation	5	calorie intake	8	goals met	12
sound	1	start-stop	8	milestones	8
		age	7	speed	7
		contact info	5	friend rank	5
		goals chosen	5	route	4
		gender	3	body mass idx	2
		blood pressure	2	daily schedule	1
		credit card	1		
		name	1		
		women's health	1		

Table 2: Number of participants (out of 30) who mentioned each category and type of data. The categories and data types emerged from the data, and were identified by the researchers as described in section 4.1.

The four most common calculated data types were distance, calories burned, activity level and sleep patterns. Participants mentioned examples of calculated data much more frequently than measured or entered data, and were aware that these data resulted from combinations of other data collected by the tracker. Fifteen out of our 30 participants used the phrase "based on" to refer to the antecedents of calculated data types.

These three categories of data types highlight ways that our participants explained and reasoned about the data collection capabilities of their activity trackers. Some of these ways of understanding how their trackers work can be learned through direct experience entering personal information. However, when it comes to measured and calculated data, our participants could only rely on indirect experiences, because they were unable to observe the more technical aspects of how the trackers collect data. For example, whereas most of our participants believed that heart rate is directly measured, in reality it is inferred using optical heart rate sensors. These sensors use LED light reflected from the skin to detect changes in blood volume as a user's capillaries expand and contract. Likewise, steps are estimated using data from accelerometers, which record force exerted by acceleration, usually in three directions simultaneously. Both of these data types rely on algorithms to identify patterns in the raw sensor data to separate out the signal from the noise and estimate when a heart beat or step took place. To do this accurately requires aggregated training data collected from many different people over time, doing different activities. However, our participants' folk theories included only a highly simplified representation of this process. In the next section we provide further explanation about how visibility is important for the development of folk theories of sensor data.

4.2 Data Visibility Supports Folk Theories

Visibility into the origins of different data types was related to the categories of beliefs about entered, measured, and calculated data. Our participants' perceptions of dependencies between data types relied on evidence that they could see and verify directly for themselves; for example, by watching step counts increase while out walking. An activity tracker's user interface itself makes some data dependencies visible, like the connection between increased step counts and calories burned or active minutes. The things that participants directly and indirectly perceived influenced their reasoning about how activity trackers collect data. This section presents findings about ways that interacting with an activity tracker device and app provides experiences that help users form folk theories about what the tracker knows, and how it works.

4.2.1 Seeing Simultaneous Changes

As users see their step counts increase at the same time as other data values in the tracker's interface, they learn which data types are related to each other. For example, steps and calories burned increase together throughout the day; therefore, each step must be connected with a certain number of calories burned. P07 described it this way: "Really like however you walked, it shows how many calories that you have burned." Steps were also spoken about as a unit of measure for other calculated data types, like distance and active minutes. When P13 was asked, "you mentioned distance, can you explain what kind of distance you mean?" she answered, "Well, it goes in steps". Three other participants also talked about how steps are somehow converted into miles for the purpose of recording distance traveled (P04, P08, P23).

Many activity trackers highlight so-called 'active minutes' in the interface in order to provide feedback to users about how active they are. P14 contrasted being "constantly on the move" with "just walking around my kitchen or something" and said that his tracker can tell the difference. If he "walk[s] quickly to and from a meeting" those minutes "[get] tallied in the active minutes section" by the tracker.

Many participants talked about ways that the tracker device or app provided feedback about how much progress they had made toward the goals they had chosen. For example, P04, P23, and P30 said that their trackers vibrate when a goal is reached. Three participants talked about how the app provided feedback in the form of a visualization. P14 described how the interface looks as it changes over the course of the day as he accumulates steps, to display progress toward his step count goal:

> And obviously once it gets to the end, I know I've had my 10,000 [steps]. And the color kind of changes, so it starts off as a dark blue and then it goes to a yellow and then a green kind of once you're really approaching your step goal.

Notifications about goals met and indicators in the user interface provide information that our participants noted as signaling cause-and-effect relationships between the data types: calories burned increases because steps are increasing, vibration was caused by meeting a goal. These relationships are what participants described to us when talking about how their trackers are able to know certain types of data.

4.2.2 Direct and Indirect Perception

Characteristics of the activity being tracked assisted users in making connections between their own experience and the data collected about the activity. This is particularly clear when contrasting steps with sleep patterns. Steps were described as "counted" or "calculated", while sleep patterns were "monitored" or "watched". Steps are a discrete action from the user's perspective that are directly experienced [13]; sleep patterns occur when the user is not awake and cannot compare what the tracker says with their own perceptions. In other words, sleep was an activity "monitored" by the tracker that participants found difficult to verify, because they could not observe the quality of their own sleep for themselves.

Participants talked about both sleep patterns and steps as based on movement. However, most participants were more skeptical about what their trackers reported about sleep patterns than steps. For example, P01 told us that she does not know how the tracker can tell she is in REM sleep, saying she doesn't know much about the "inner workings" of the tracker, but that sleep tracking involves "more variables than what you can sense on your wrist". P09 said that her "completely unscientific theory" is that the tracker can tell how well she slept because it can tell when her arm stops moving and when her heart rate drops. She was satisfied with this explanation because, "it has never accidentally thought I was sleeping". In other words, she was confident in the behavior of the tracker while she was asleep, because the experience she had with it when she was awake led her to believe that it was working properly. However, P05 had the opposite experience. She had observed her tracker making the error that P09 said had never happened to her:

> But a lot of times it's not necessarily registering I'm awake, when I have a kid whose head is right on top of me and I refuse to move or something, and I'm just laying there for an hour. Like, it's not... So, I think it's doing movements.

There were in fact two participants who talked about tracking sleep patterns more confidently. These participants had less sophisticated trackers that required them to manually start sleep mode each night if they wanted to track sleep patterns. For example, P24 said,

> When I go to sleep, I have to double tap it and then it records how many hours I sleep, and it also records my movement in my sleep, so that it shows me when I'm in deep sleep because I won't be moving, and when I'm restless throughout the night it shows on a little graph.

These examples illustrate that participants were most confident about the aspects of sleep tracking that happened while they were still awake and could connect their perceptions to what the tracker was reporting, and the least confident about the aspects that they could not observe while they were asleep. Being able to connect the state transition from tracking "movement" to tracking "sleep" with something they could perceive directly, witness, or enact themselves (like seeing it change when they were very still, or manually turning on sleep mode) made sleep tracking seem more believable.

4.2.3 Visibility of the Sensor

In talking about heart rate data collection, participants were focused on the sensor—where it is worn, what it is doing, and how one can see the sensor readings. Heart rate was more strongly connected with the concept of a sensor than any of the other measured data types. Having one's heart rate measured during a visit to the doctor is a common experience, and heart rate has a medical and fitness interpretation that many people are already aware of. Many participants referred to the physical part of the body where a tracker with a heart rate sensor should be worn, which, as P09 described, is worn "a little bit above your wrist, and it has a little sensor, it's like a green light actually". P16 explained how her tracker is able to know her heart rate:

> The one that I have, the Charge HR, it measures your heart rate based on your wrist. There's a sensor that I

> don't know the specifics of, that you wear one finger away from your wrist. So it's tracking your heart rate there.

Twenty four participants mentioned heart rate, but only nine participants said they used a tracker with a heart rate sensor. For example, P04 used a Fitbit wristband that does not have heart rate capability, but was aware that some trackers can do this:

> But I do know, on the wrist, that some of them can track your heart rate because obviously, that's where... If you're going to the doctor's office, that's where they're checking your pulse at.

Heart rate and movement data present an interesting point of comparison regarding participants' folk theories. Far fewer participants mentioned movement than heart rate (17 versus 24). Heart rate sensors are visible components of the device, because they must be on the outside of the device to work properly. In contrast, accelerometers, which collect movement data, are inside the device and cannot be seen. If movement was mentioned, it typically only came up as a way of explaining how the tracker was able to detect other data types like sleep patterns, flights of stairs or number of steps. For example, P27 said he was not sure about how the tracker could tell how many flights of stairs he has gone up, but said there's a "motion sensor for kind of the jarring sensation that would be given by going a single step". P09 also mentioned the sensor in relation to flights of stairs, and how she was unsure what kind of sensor allowed the device to have this capability: "And so it's obviously some kind of sensor that's just not in the other equipment [her previous tracker]."

It is as though these participants only knew about the "motion" sensor because they were trying to reverse-engineer where the calculated data values came from. However, being able to actually see a part of the heart rate sensor component (e.g., the green LED on the Fitbit Charge HR), or having to intentionally interact with it to take a measurement (put a finger on the flash, as P22 talked about with her Samsung Galaxy smartphone) makes the sensor itself more salient, making the data generated by the sensor more salient as well. This difference in visibility of the sensor and the perceptions about data provenance that visibility enabled was an important differentiator for our participants between measured and calculated data types.

4.2.4 Perceived Inaccuracy

Seventeen participants described noticing that their tracker counted an activity differently than they expected. For example, P23 noticed a discrepancy while applauding at a show:

> I was at a show and I clapped and I saw that [the tracker] was lighting up and then a friend of mine, who I was there with, he had a fancier, I have just the one that has lights, but his tracker actually, you could press the button and see how many steps it was. And so then after the next song we clapped again, we looked before we clapped the number of steps he had and then he clapped, and then he looked again and it was higher.

This anecdote illustrates an observation made by other researchers [13, 43, 56] who have written about the experimentation that activity tracker users engage in when they notice perceived inaccuracies in their data.

The physical display on the wrist-worn Fitbit trackers is limited, and can only display one piece of information at a time, typically a count of a data type like steps, stairs climbed or calories burned. People who wear trackers on their wrists, rather than in their pockets or elsewhere, have more opportunities to notice the disparity between their perceptions of movement and the tracker's step counts. This is because the tracker's display is more visible when worn on the wrist. These participants talked about how noticing this disagreement inspired them to assess the accuracy of their tracker's performance, and to speculate about how the calculated data are produced. For example, P03 said that when she pushes the stroller, she thinks the tracker is not "calculating" because she's "not getting any steps." P20 also made a very similar comment: she said that her tracker underestimates her activity when she is pushing the stroller or holding the dog's leash, because she isn't swinging her arm back and forth as much when doing those activities.

Perceived inaccuracies made visible by the device's display encouraged speculations about discrepancies between how the tracker works and participants' subjective perceptions of their movements. Experiencing these discrepancies provided opportunities for our participants to incorporate additional information into their folk theories about how their trackers collect data.

4.2.5 Manual "Recording"

Participants described using an input mechanism provided by the tracker to enter information about the beginning and end of periods of time taken up by certain kinds of activities, such as exercise or sleep. By entering this data, users can mark a change from one state of activity to another. These state changes indicating when activities start and end add a layer of context to a particular timeframe, in which the tracker then uses the sensor data it collects to determine active minutes or sleep rather than steps.

For example, many trackers offer users the ability to manually log duration and type of activity. Similar to the automatically detected active minutes, entering this information changes how the device interprets data recorded during that time period from inactive, to active. P15 talked about pressing the "record" button to enter a mode that tracks "how far you went, and the calories you personally burned, if you're hiking":

> And then, there's a little record button just like you would have on your voice recording or whatever, if you're recording a video or whatever. And then you just press it to stop which is pretty neat. (P15)

P28 talked about something similar, regarding turning on and off the GPS so that:

> ...it'll mark how far you've walked. And then when you tell it to stop, it's like, 'Okay, well you've walked one mile at this pace so you burned this many calories.'

In the above instances, the user provided information to the tracker that marked a state transition from one category of activity to another, enabling her to see a representation of the data calculated by the tracker in the interface that matched her own awareness of and intention for what she was doing at that time.

In addition to specifying periods of higher activity, some trackers allow users to manually specify that they have entered sleep mode, which changes the tracker's interpretation of movements registered by the device from steps to restless sleep. P01, P05, P12, P24, and P25 all described how they manually "double tap" the tracker (or "hit it twice really fast") to make it enter sleep mode. P01 said, "You tell it when you go to sleep and you tell it when you wake up and it tracks how you were sleeping."

However, as, P05 described, this manual stopping and starting feature has some limitations:

> So for sleep you have to set it, like you have to tell it. So I'm inputting that. I'm turning it into sleep mode. I'm turning it off of sleep mode. Although I forgot to turn it off today till like one o'clock. But [laughter]... No, I did not sleep till one o'clock unfortunately, I wish.

This form of manual data entry enables the device to collect a different kind of data for that time period, like a higher activity level or calorie usage. Using an input mechanism to tell the tracker that certain data should be interpreted as being related to a particular activity helped participants to become more aware of what kinds of data the tracker can and cannot collect, and when. By thinking about data collection as something that must be started and stopped, like turning on a recording device, it supports a more limited set of expectations about what data the tracker can collect on its own.

5. DISCUSSION

Users are expected to self-manage their privacy by making choices about consent for what kinds of data collection to allow. However, they cannot do this effectively if they cannot reason about what kinds of data collection and inferences are possible. Our study focuses on folk theories, because this allows us to understand how activity tracker users think about their data, and therefore what their knowledge and experience allows them to base their privacy-related decisions on. Knowing more about their folk theories can help us better understand the boundaries between what users can and cannot reasonably consent to. Our design implications suggest ways to encourage speculation and thereby broaden users' folk theories, which could help them to better self-manage their privacy.

Our findings indicate that folk theories of activity tracker data collection arise from information provided in the interface, and from users' own perceptions of their activities. The folk theories we elicited involved three categories of data: that which users *enter* about themselves like age and weight, data that are *measured* by the tracker like steps and location, and data that are *calculated* based on other data like activity level, distance and calories burned. However, these folk theories about data types do not include other kinds of information that might be inferred from the raw sensor data generated by activity trackers, but are not directly related to activity tracking. In other words, the folk theories are constrained by what participants use activity trackers for.

Conceptualizing steps as a discrete unit of measurement, for example, supports reasoning about physical activity and fitness. But at the same time, it prevents understanding that in order to identify a step the tracker must engage in a statistical classification task. It also prevents the realization that if movement data can be used to count steps, other movements the tracker detects could be used to count other kinds of actions. This means that activity tracker users whose folk theories do not include movement as a measured data type or who do not know that steps and sleep are estimates based on movement are unlikely to be able to truly consent to the collection of data types that are calculated based on movement data. Even a belief that both steps and sleep patterns, two very different kinds of activities, are based on movement did not inspire our participants to speculate about other kinds of data that might be derived from movement.

While our participants were inspired to speculate about some aspects of the collection of certain types of activity data, there seem to be few opportunities presented by activity trackers for users to engage in the kind of speculative reasoning that generalizing beyond what the tracker was directly presenting to them would require. For example, no one who mentioned GPS, location or distance said that their tracker knows where they live, either as part of the initial free list activity or during the follow up questions and probing. This poses a problem, from a privacy perspective, for users considering whether to consent to sensor-related data collection: if users' folk theories do not include a framework for reasoning about possible inferences from sensor data, they cannot make informed choices about which systems to use and what information they do and do not want collected about them. However, our findings point to ways that interfaces might be designed to induce the kind of speculation and thinking that would engender the development of folk theories that would be more helpful for privacy-related consent decisions.

5.1 Revealing the Context of Production

Activity tracker systems involve sensor technologies, devices, apps, and cloud services that all play a part in transforming the raw sensor data into information representing actions (e.g., steps) and physiological processes (e.g., heart rate) that users can see and understand. One important input into the folk theories of participants in our study was experiences they'd had that provided visibility into how data are produced, such as seeing step counts, active minutes, and calories burned increase together in the interface, or noticing inaccuracies. However, knowing there was a heart rate sensor and seeing their heart rate in the interface did not help users in our study to become more aware of how the device is able to determine their heart rate. For example, only three participants talked about math (P15, P21) formulas (P21), or in one case, an algorithm (P23) operating on data that their trackers collected.

Raw data, or the direct output from the sensors in the activity trackers, does not have meaning by itself. It only becomes meaningful after being processed and presented to the user, in such a way that they can see themselves in their data [49]. This transformation is work that the system does on the user's behalf, so that they do not have to track their activities and perform those calculations themselves. The interface between the user and tracker hides this work, so that users are given no cause consider that step counts are not raw data. For the activity tracker users in our study, raw sensor data was not a salient aspect of *information that an activity tracker knows*. This hidden work presents a challenge for supporting folk theory development; because folk theories arise from people's experiences, users must be able to encounter or experience some aspect or evidence of this work for it to be incorporated into their folk theories.

Vertesi et al. [50] wrote about the importance of knowing the context of data production, or "how the data is crafted and acquired," in scientific collaborations. They emphasized that hiding the work that goes into preparing scientific data for sharing outside the team that produced it obscures the sociotechnical infrastructure that gives it value and meaning. In an activity tracker system, sensors, raw data, processing and other infrastructure are also invisible to the users who interact with the final output in the displays of their activity tracker devices and apps. In packaging up raw sensor data as activity data, the details of the context of production are left out in order to allow the activity data to gain credibility, resulting in processed data that seem more definitive and "true" than they really are. In other words, hiding the relationship between what the system is doing and what the user sees prevents the user from developing folk theories about data as interpretations and inferences, not absolute facts. Obscuring the ambiguity may help people become more confident in the data, but it also prevents them from speculating about what else it might be used to infer, and forming folk theories that incorporate ideas about data processing, transformation and dependencies. Information that is not incorporated into people's folk theories cannot help them to imagine potential consequences of data collection, or reason about privacy-related decisions.

5.2 Implications for Design

The seamless approach to the design of ubiquitous computing systems, as Weiser said, "focusing on the task, not the tool" [54], hides uncertainty by replacing it with certainty [9]. However, Kay and Kummerfeld [20] argue that systems should be *scrutable*, or understandable through study and observation. A scrutable system has an interface that allows the user to see the "evidence source" and the "interpretation processes" that produce the information that is consumed. Bellotti and Sellen [5], in an early paper about designing for privacy in ubiquitous computing systems, wrote about empowering users by creating designs that provide feedback about these invisible aspects. It may therefore be better for privacy to be less seamless and more scrutable; to look for ways to reveal hidden work and help users make connections between the data collection and dependencies they are already aware of in the activity tracking context, and other information that may be only indirectly related to that context.

One challenge inherent to making the production of activity tracker data more observable is that users may find the additional information overwhelming and not know what to do with it. For example, Rapp and Cena [39] found that people who had never used activity trackers before participating in their study felt the data and graphs the trackers provided were already "too abstract and removed from what they were expecting", not meaningful to them, and difficult to engage with. However, our findings suggest several ways that small design changes to the information provided in the tracker's interface might support the development of folk theories through encouraging speculation about how the data are produced.

Seeing simultaneous changes to multiple data types in the app interface (e.g., steps and calories burned) led to folk theories that incorporated causal relationships between those data types. But, participants needed a reason to be looking at the interface in the first place in order to see the relationship between those data types, and that reason is activity tracking. Presenting information about other kinds of data dependencies that are related to but not directly about activity tracking may be a minor departure from the user's main task that engenders speculation about what else an activity tracker might know.

Many services based on sensor data periodically publish essays on the company's blog or website providing analysis of patterns in the data generated by users of the service; Fitbit is one example of this[4]. If activity tracker service providers were to incorporate information comparing users' data with aggregate statistics as part of the app's interface, it could provide additional visibility into the aggregation that underlies all of the data output that users interact with. For example, when reporting sleep patterns, the app could also present information like, "Your average bed time is 11:23 PM, which is 20 minutes later than other users in your age group." Alternately, to promote awareness of the possibility that a user's location might be used to generate new data about semantic aspects of geography such as where the user lives, the tracker could display to the user information about how far the participant went from home that day while jogging (not just length of the run), or how far from home their number of steps that day would have taken them. In a more "creepy" vein [44], an activity tracker app might inform users that "people who have restless nights that are similar to yours are likely to be new parents." Folk theories incorporating the kinds of insights that can be derived through aggregation might allow users to consider consequences like this when reasoning about possible privacy-related effects of using sensor-enabled technologies.

Tracking an activity that users can't directly perceive, like sleep, led to doubt and speculation from our participants about how the tracker could measure a phenomena like this. Sleep is unique in the context of activity tracking, in that it is the only activity that is not verifiable by the user while it is happening. However, other kinds of activities that might be detected also have this characteristic, to varying degrees. For example, Fitbit trackers began providing information about "stationary time", or amount of time spent without moving in a given time interval, to users in April 2016, after our study was conducted. It may be difficult for users to pay attention to the absence of an activity, but trackers can do this easily. It therefore might be possible to combine information about stationary time with GPS, and highlight data types in the interface like time spent sitting at work, or in a moving vehicle. Making these data visible could encourage users to think about how the tracker defines "stationary time," how the data are collected, how the location categories are defined, and how different data types can be combined to produce new data.

In tracker devices with an optical heart rate sensor, *visibility of the sensor component* made the source of the data collected by the device more salient for our participants, and changed the way they reasoned about the data. With the current trend towards making trackers look less like fitness equipment and more like clothing accessories, making additional sensors more visible seems like an unlikely possibility. However, perhaps there is a way to make the raw data more conceptually tangible. It might be possible to quantify aspects of the tracker itself, like the tracker quantifies aspects of the person. Many personal computers include widgets and control panels that present statistics about the "health" of the device, such as available memory, temperature, fan speed, and uptime. Similar kinds of data could be calculated about the tracker device, or about user interactions with the tracker. For example, data about how many times the user has checked the tracker's display in the last week might make the device more salient to the user in ways that are both informative and provide a focus on technical details for users to speculate about and incorporate into their folk theories.

As others have found [13, 56], *perceived inaccuracy* prompts attention to aspects of how the data are collected. It may also present a view into the statistical model that data like step counts are based on. This was a powerful mechanism supporting speculation among the users in our study about how their trackers counted steps and measured sleep. However, this speculation only extended to data types they knew the tracker was supposed to be collecting. Perceived inaccuracy highlights uncertainty in the underlying machine

[4] `https://blog.fitbit.com/how-do-your-sleep-habits-stack-up/` (accessed on July 10, 2017)

learning models, and therefore is a direct way to encourage users to notice and think about the context of data production. Folk theories that incorporate concepts related to the production of data may help users to reason about inferences and calculated data types.

Consolvo et al. [10] wrote that it is important for future research to consider better ways to present uncertainty to the user, and to understand its effects on user behavior. However, they also said that this is challenging, because typical ways of presenting statistical uncertainty are unlikely to be understood by most users. The challenge for design to support folk theories of data collection is how to provide information that helps the user connect the realization that a tracker may be collecting some kind of data other than steps, to specific other kinds of information or activities the tracker might be able to detect. "Glanceable" displays on activity trackers with a smart watch form factor traditionally have been focused on presenting status updates related to activity over the past hour, goal attainment, etc. [15]. But it might be possible to use the displays to notify users about some of the uncertainty involved in activity recognition, by using colors or shapes to indicate deviation from the underlying statistical models.

Finally, *manual "recording" of activities* via state transitions that are entered by the user, like activating sleep mode, are also a form of data collection. These data give the tracker additional context to use to interpret the raw sensor data collected during certain time periods. The act of starting and stopping the "recording" also gives the user more confidence in the accuracy of the data that are collected about the activity. If the user were able to provide other kinds of contextual information to the tracker, it could both help improve the functioning of the system, and also help the users better understand the context of production. For example, activity tracker users could be given the ability to contribute data consisting of feedback on the tracker's performance. A "thumbs up" or "thumbs down" might signal points at which they feel the tracker is particularly accurate or inaccurate. Data like this collected over time and relayed back to the user in aggregate might provide visibility into the messiness of the context of production and the work that goes into estimating step counts, while also providing information that users would find helpful for understanding when they can trust the tracker and when they cannot, and that system operators would find helpful for improving accuracy.

5.3 Implications for Privacy Self-Management

Design to encourage speculation about the context of production of activity tracker data has implications for the formation of folk theories about sensor data collection, and for helping users make decisions about privacy self-management and consent. Folk theories are "ways of understanding" [12] that are based on experience and help users of technologies make decisions [53]. In other words, folk theories are cognitive structures that help users envision what might happen based on what they already know. A folk theory that includes knowledge about sensors and the kinds of data an activity tracker records about the world, or the concept that the numbers displayed by the tracker are estimates with a degree of uncertainty, or that some data are produced by combining other types of data, may help users to speculate and imagine different possible consequences than a folk theory that involves certainty that steps are directly counted.

This does not mean that folk theories need to be technically accurate from an expert's perspective. Kempton demonstrated in his thermostat study that incorrect mental models about technology can still be useful for decision-making; in his case, for making home heating decisions [22]. It is not necessary for a user to understand how an accelerometer works, or what the algorithm for identifying a step is, to speculate that if step counts are estimates other kinds of information may be estimated too. Speculation does not need to produce accurate knowledge to be useful for reasoning about control over data collection and possible consequences. Folk theories that do not help users reason about possible consequences beyond health and fitness may not be helpful for making consent choices about data collection in systems that involve inferences beyond the direct context of use. Folk theories that involve speculation about aspects of the context of production could provide better support for informed privacy self-management and consent.

6. CONCLUSION

Sensor-enabled systems, like activity trackers, collect highly detailed and personal data about users' behavior. Because people are expected to be able to self-manage their privacy regarding digital information, it is important to understand users' folk theories of this sensor data collection, which help them reason about new situations and make decisions. Our findings show that users' folk theories are limited to the activity tracking context, and do not help users reason about other kinds of data that might be collected or used beyond activity tracking. Instead, activity tracker interfaces obscure the complexity and uncertainty involved with producing the data that are shown to users. By hiding the messiness of transforming raw data into useful insights, the data that are collected become more helpful for the user's primary task (health and fitness), but not useful for reasoning about privacy, which is at best a background task.

Despite this, users have experiences with their trackers that open them up to speculating about how their data are produced, and to learning about connections between data types. While designs that provide hints about some of the complexity may come with some cost for the user, our findings suggest avenues for design that build on speculation users are already engaged in, in ways that are peripherally related to current tracker functionality. Future work is needed to further understand the connection between speculation, folk theories about data collection, and user reasoning about privacy and consent.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1524296.

8. REFERENCES

[1] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal. Your Location has been Shared 5,398 Times! In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 787–796, 2015.

[2] F. Asgharpour, D. Liu, and L. J. Camp. Mental Models of Security Risks. In *International Conference on Financial Cryptography and Data Security*, pages 367–377, 2007.

[3] R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. "Little brothers watching you:" raising awareness of data leaks on smartphones. In *Symposium on Usable Privacy and Security*, page Article 12, 2013.

[4] P. Bayman and R. E. Mayer. Instructional Manipulation of Users Mental Models for Electronic Calculators. *International Journal of Man-Machine Studies*, 20(2):189–199, 1984.

[5] V. Bellotti and A. Sellen. Design for Privacy in Ubiquitous Computing Environments. In *Proceedings of the European*

Conference on Computer-Supported Cooperative Work, pages 77–92, 1993.

[6] H. R. Bernard and G. Ryan. *Analyzing Qualitative Data: Systematic Approaches*. SAGE Publications, 1st edition, 2009.

[7] S. P. Borgatti. Elicitation techniques for cultural domain analysis. In J. Schensul and M. LeCompte, editors, *The Ethnographer's Toolkit, Vol. 3*, pages 115–151. Altimira Press, 1998.

[8] D. D. Brewer. Supplementary Interviewing Techniques to Maximize Output in Free Listing Tasks. *Field Methods*, 14(1):108–118, 2002.

[9] M. Chalmers, I. MacColl, and M. Bell. Seamful design: Showing the seams in wearable computing. In *Eurowearable*, pages 11–16, 2003.

[10] S. Consolvo, P. Klasnja, D. W. McDonald, and J. A. Landay. Designing for Healthy Lifestyles: Design Considerations for Mobile Technologies to Encourage Consumer Health and Wellness. *Foundations and Trends in Human-Computer Interaction*, 6(3-4):167–315, 2014.

[11] M. Eslami, K. Karahalios, C. Sandvig, K. Vaccaro, A. Rickman, K. Hamilton, and A. Krilik. First I Like It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2371–2382, 2016.

[12] S. A. Gelman and C. H. Legare. Concepts and Folk Theories. *Annual Review of Anthropology*, 40:379–398, 2011.

[13] N. Gorm and I. Shklovski. Steps, Choices and Moral Accounting: Observations from a Step-Counting Campaign in the Workplace. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 148–159, 2016.

[14] S. D. Gosling, S. Vazire, S. Srivastava, and O. P. John. Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, 59(2):93–104, 2004.

[15] R. Gouveia, F. Pereira, E. Karapanos, S. A. Munson, and M. Hassenzahl. Exploring the design space of glanceable feedback for physical activity trackers. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 144–155, 2016.

[16] B. Hulkower. Fitness Fuels US Consumer Desire for Wearable Tech, 2016. `http://www.mintel.com/press-centre/technology-press-centre/fitness-fuels-us-consumer-desire-for-wearable-tech-with-sales-increasing-186-in-2015`.

[17] A. Jackson. Husband and Wife Never Expected Their Fitbit Would Tell Them This, Feb. 2016. `http://www.cnn.com/2016/02/10/health/fitbit-reddit-pregnancy-irpt/`.

[18] P. N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

[19] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Symposium on Usable Privacy and Security*, pages 39–52, 2015.

[20] J. Kay and B. Kummerfeld. Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems*, 2(4):Article 24, 2012.

[21] F. C. Keil. The Feasibility of Folk Science. *Cognitive Science*, 34(5):826–862, 2010.

[22] W. Kempton. Two Theories of Home Heat Control. *Cognitive Science: A Multidisciplinary Journal*, 10(1):75–90, 1986.

[23] P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower. Exploring Privacy Concerns about Personal Sensing. In *International Conference on Pervasive Computing*, pages 176–183, Mar. 2009.

[24] J. Lin, S. Amini, J. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the ACM Conference on Ubiquitous Computing*, pages 501–510, 2012.

[25] T. R. Lindlof and B. C. Taylor. *Qualitative Communication Research Methods*. Sage Publications, 2nd edition, 2002.

[26] T. Matlock, S. C. Castro, M. Fleming, T. M. Gann, and P. P. Maglio. Spatial Metaphors of Web Use. *Spatial Cognition and Computation*, 14(4):306–320, 2014.

[27] J. A. Maxwell. *Qualitative Research Design: An Interactive Approach*, volume 41 of *Applied Social Research Methods*. Sage Publications, 3rd edition, 2012.

[28] M. B. Miles and M. A. Huberman. *Qualitative Data Analysis: An Expanded Sourcebook*. Sage Publications, 2nd edition, 1994.

[29] V. G. Motti and K. Caine. Users' Privacy Concerns About Wearables. In *International Conference on Financial Cryptography and Data Security*, pages 231–244. 2015.

[30] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[31] G. Novick. Is there a bias against telephone interviews in qualitative research? *Research in Nursing & Health*, 31(4):391–398, 2008.

[32] G. Paolacci, J. Chandler, and L. N. Stern. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[33] S. Petronio. *Boundaries of Privacy: Dialectics of Disclosure*. State University of New York Press, 2002.

[34] E. S. Poole, M. Chetty, R. E. Grinter, and W. K. Edwards. More than meets the eye: Transforming the User Experience of Home Network Management. In *ACM Conference on Designing Interactive Systems*, pages 455–464, 2008.

[35] E. S. Poole, C. A. Le Dantec, J. R. Eagan, and W. K. Edwards. Reflecting on the Invisible: Understanding End-User Perceptions of Ubiquitous Computing. In *Proceedings of the International Conference on Ubiquitous Computing*, pages 192–201, 2008.

[36] M. Quinlan. Considerations for Collecting Freelists in the Field: Examples from Ethobotany. *Field Methods*, 17(3):219–234, 2005.

[37] E. Rader and R. Gray. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 173–182, 2015.

[38] A. Raij, A. Ghosh, S. Kumar, and M. Srivastava. Privacy Risks Emerging from the Adoption of Innocuous Wearable Sensors in the Mobile Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20, 2011.

[39] A. Rapp and F. Cena. Personal Informatics for Everyday Life: How Users Without Prior Self-Tracking Experience

Engage with Personal Data. *International Journal of Human-Computer Studies*, 94:1–17, 2016.

[40] A. K. Romney, D. D. Brewer, and W. H. Batchelder. Prediciting Clustering From Semantic Structure. *Psychological Science*, 4(1):28–34, 1993.

[41] T. A. Schwandt. *Qualitative Inquiry: A Dictionary of Terms*. Sage Publications, 1997.

[42] F. Shih, I. Liccardi, and D. Weitzner. Privacy Tipping Points in Smartphones Privacy Preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 807–816, 2015.

[43] P. C. Shih, K. Han, E. S. Poole, M. B. Rosson, and J. M. Carroll. Use and Adoption Challenges of Wearable Activity Trackers. In *iConference Proceedings*, 2015.

[44] I. Shklovski, S. D. Mainwaring, H. H. Skúladóttir, and H. Borgthorsson. Leakiness and Creepiness in App Space: Perceptions of Privacy and Mobile app Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2347–2356, 2014.

[45] D. J. Solove. Privacy self-management and the consent dilemma. *Harvard Law Review*, 126(7):1880–1903, 2013.

[46] J. E. Sturges and K. J. Hanrahan. Comparing telephone and face-to-face qualitative interviewing: a research note. *Qualitative Research*, 4(1):107–118, 2004.

[47] The White House. Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy. *Journal of Privacy and Confidentiality*, 4(2):Article 5, 2012.

[48] E. C. Thompson and Z. Juan. Comparative Cultural Salience: Measures Using Free-List Data. *Field Methods*, 18(4):398–412, 2006.

[49] P. Tolmie, A. Crabtree, T. Rodden, J. A. Colley, and E. A. Luger. "This has to be the cats:" Personal Data Legibility in Networked Sensing Systems. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 490–501, 2016.

[50] J. Vertesi and P. Dourish. The Value of Data: Considering the Context of Production in Data Economies. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 533–542, 2011.

[51] R. Want, A. Hopper, V. Falcão, and J. Gibbons. The Active Badge Location System. *ACM Trans. Inf. Syst.*, 10(1):91–102, 1992.

[52] W. H. Ware and the Secretary's Advisory Committee on Automated Personal Data Systems. *Records, Computers, and the Rights of Citizens*. US Department of Health, Education & Welfare, 1973. `https://www.justice.gov/opcl/docs/rec-com-rights.pdf`.

[53] R. Wash. Folk Models of Home Computer Security. In *Symposium on Usable Privacy and Security*, page Article 11, 2010.

[54] M. Weiser. The world is not a desktop. *interactions*, 1(1):7–8, 1994.

[55] S. C. Weller and A. K. Romney. *Systematic Data Collection*. Qualitative Research Methods Series 10. Sage Publications, 1st edition, 1988.

[56] R. Yang, E. Shin, M. W. Newman, and M. S. Ackerman. When Fitness Trackers Don't 'Fit': End-User Difficulties in the Assessment of Personal Tracking Device Accuracy. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 623–634, 2015.

[57] Y. Yao, D. Lo Re, and Y. Wang. Folk Models of Online Behavioral Advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1957–1969, 2017.