Models Matter: Setting Accurate Privacy Expectations for Local and Central Differential Privacy

Mary Anne Smart
Purdue University
masmart@purdue.edu

Priyanka Nanayakkara
Harvard University
priyankan@g.harvard.edu

Rachel Cummings
Columbia University
rac2239@columbia.edu

Gabriel Kaptchuk
University of Maryland, College Park
kaptchuk@umd.edu

Elissa M. Redmiles
Georgetown University
elissa.redmiles@georgetown.edu

Abstract

Differential privacy is a popular privacy-enhancing technology that has been deployed both by industry and government agencies. Unfortunately, existing explanations of differential privacy fail to set accurate privacy expectations for data subjects, which depend on the choice of deployment model. We design and evaluate new explanations of differential privacy for the local and central models, drawing inspiration from prior work explaining other privacy-enhancing technologies such as encryption. We reflect on the challenges in evaluating explanations and on the tradeoffs between qualitative and quantitative evaluation strategies. These reflections offer guidance for other researchers seeking to design and evaluate explanations of privacy-enhancing technologies.

Keywords

differential privacy, human factors, usable security and privacy

1 Introduction

A core focus of usable security and privacy research is making a system's privacy protections transparent to end users in order to facilitate informed decision-making about data sharing [29, 32, 63]. One such privacy protection is differential privacy (DP). DP [28] is a privacy-enhancing technology that has been rapidly adopted by industry and government agencies [2, 22, 30, 70, 71, 96]. DP deployments provide provable privacy guarantees by adding statistical noise to computations; this noise obfuscates the information of each individual while preserving aggregate-level insights. In response to DP's rapid success, a growing body of work has started to document the inadequacies of existing messaging around DP [18] and design new explanations for DP systems (e.g., [12, 34, 50, 76, 91, 106, 107]; see [24] for a survey of this work). In particular, prior work has found that existing descriptions of DP fail to articulate critical deployment information that is necessary to understand the privacy guarantees [18]. In this work, we leverage effective techniques from the usable security literature to explain DP and explore the value of both qualitative and quantitative evaluation in assessing the efficacy of using these techniques in comparison to the state-of-the-art.

Specifically, we develop messaging for DP that highlights the threat models that are implicit in different approaches to deploying DP. These implicit threat models are critical for end users to understand before sharing their data, as the chosen threat model may not provide protections against the classes of attackers about which they are concerned. We do this by exploring three explanation formats drawn from the existing PETs messaging literature: nutrition labels [52], diagrams [91, 93, 107], and metaphors [21, 50, 82, 94, 109, 110]. Each of our evaluated explanations aims to communicate the consequences of DP in terms of which information flows the deployment protects against.

Differential Privacy Deployment Models. There are multiple deployment *models* for DP, each of which is associated with a particular threat model. The two most widely deployed models are the *central model* [28] and the *local model* [51].[1]

The central model assumes there exists a data curator who is *trusted* to see raw data from individuals; the adversary can only access released, aggregate results. The data curator collects data from individuals, performs statistical analyses on the collected dataset, and then injects statistical noise into the results before release. This process limits the ability of the adversary to learn about individual records from summary statistics, at the cost of reduced accuracy. The potential danger of this model, however, is that the data curator might *not* be trustworthy; the database storing individuals' data could be vulnerable to hackers or misuse by insiders if other, complementary security practices are not adopted in tandem. Well-known deployments of the central model include the U.S. Census Bureau's data products for the 2020 Decennial Census [2].

In the local model, noise is added to each individual's data *before* collection, meaning the unmodified data are never stored together. As such, there is not the same need to trust the data curator (i.e., it is assumed that the data curator is honest but curious). This higher level of security comes at a cost: significantly more noise must be added to the data in order to ensure the same level of privacy protection, reducing the accuracy—and, thus, utility—of the collected data. Notably, Google and Apple have both used local DP to analyze browser data in Chrome and Safari, respectively [4, 30].

Helping Users Understand DP Models. Ensuring that descriptions of DP accurately convey information about the model is crucial to designing transparent messaging for DP deployments: the threat surface associated with the two main models differ significantly,

[1] Although there are many variations of DP [23], we focus on the *central* and *local* models due to their popularity.

Proceedings on Privacy Enhancing Technologies 2025(4), 653–678

https://doi.org/10.56553/popets-2025-0150

even if they provide the same privacy guarantees for the *data releases*. Specifically, data collected under the central model can be hacked, leaked, or abused by an insider threat if sufficient complementary privacy measures are not taken, while data collected under the local model does not share these risks.

Data subjects cannot be expected to make informed data-sharing decisions if they believe that DP is "some sort of crypto-magic to protect people from data misuse" [87]. Prior work has demonstrated that existing DP description strategies do a poor job aligning users' privacy expectations with the privacy protections provided by different DP models [18]. In other words, the kind of protection that users expect does not align with the actual nature of the protection offered by DP. Misaligned expectations exacerbated by poor communication can lead to data subjects underestimating or—even more alarmingly—overestimating the privacy protection that DP offers. In our work, we use a mixed-methods approach to explore methods of explaining the implications of the deployment model. Rather than explaining how DP works from a technical perspective, we focus primarily on helping people understand which threats are prevented by the local versus central models. We want descriptions of DP that set different privacy expectations depending on whether the local or central model is used. Since our DP descriptions focus on implications rather than technical details, we expect them to be adaptable to explaining implications of other PETs with similar goals.

We first explore three kinds of explanations that build on best practice from usable security [18, 50, 52, 77, 82, 107]: metaphors, diagrams, and privacy labels for DP. Following the methodology adopted in prior work [31, 90, 101], we conduct this exploration in two phases: (1) we start with qualitative methods to explore the design space, and (2) follow up with a quantitative evaluation. Specifically, we begin with an interview study through which we identify the most promising strategies—privacy labels and metaphors—and further refine these explanations based on participant feedback. Then, we evaluate our refined explanations in an online survey (n = 698), measuring objective comprehension, subjective understanding, perceived thoroughness, and trust. We compare our explanations to each other and text-based explanations of DP [106]. Our results offer insight for future work explaining PETs. For example, we show promising evidence that adapting the idea of nutrition labels for privacy to explain DP can support reasoning about which data flows are protected. We also show both qualitatively and quantitatively the importance of information about the mechanism of privacy protection, not just the implications of that protection, on people's perceived confidence in their understanding of how their data is protected.

The process of investigating design strategies also allowed us to characterize the mental models people form around DP. While studying mental models was not the primary goal of our study, we discuss particularly interesting insights that can guide future work. For example, we found that participants often tried to make sense of DP through comparisons to other PETs such as encryption. We conclude with a discussion of the potential design implications of our findings.

Finally, we reflect on the challenges of evaluating explanations of privacy-enhancing technologies, viewing our work on explaining DP as a case study. One challenge lies in understanding the trade-offs between qualitative and quantitative evaluation methods and determining an appropriate balance between qualitative and quantitative evaluation. Perhaps the greatest challenge lies in defining what it means for an explanation to be effective and selecting the relevant metrics. While prior work has proposed a range of evaluation metrics, there does not yet exist a widely-accepted standard. We suggest that developing such a standard, and choosing the appropriate metrics, requires clear articulation of an explanation's goals.

2 Background

A growing body of work provides guidance for effective security and privacy (S&P) communication [36, 88]. Awkward interfaces or ineffective communication can lead to dangerous misconceptions and risky behaviors [18, 39, 58, 104]. One reason that people may misjudge privacy risks or misuse PETs is that they lack appropriate *mental models*. Prior work has argued that "efficacy of risk communication depends not only on the nature of the risk, but also on the alignment between the conceptual model embedded in the risk communication and the user's mental model of the risk" [5]. Unfortunately, existing depictions of DP appear to be misaligned with people's mental models, resulting in misaligned privacy expectations [18].

In this section, we outline the relevant prior work on the challenges of designing effective, transparent communication about PETs. First, we discuss the prior work on communicating with data subjects about DP. Next, we discuss three particularly popular privacy explanation strategies—metaphors, diagrams, and nutrition labels. Finally, we discuss prior work on mental models in S&P.

Implications vs Process. One of the most important findings from prior work is that explaining data protection processes is not enough for most readers to grasp the implications of the protection offered by PETs [18, 26, 91, 106]. Xiong et al. [106] studied explanations of both central and local DP, and found that when the implications of the local and central models were stated explicitly, participants were more willing to share information under the local model. Kühtreiber et al. [59] replicated this study with German participants, finding that the participants who read descriptions of local DP that explained its implications explicitly were more willing to share personal information than any other group. Unlike these studies, we explore a variety of best-practice methods from the usable S&P literature (i.e., metaphors, diagrams, and nutrition labels) to communicate which information flows are protected by DP under the central and local models. Our focus on communicating information flows about which people care distinguishes our study from prior work, in which only a few information flows are typically explored.

Cummings et al. [18] explored the implications of DP through six information disclosures about which people care and against which DP may protect—depending on whether the local or central model is used. They found that existing descriptions of DP fail to appropriately set privacy expectations regarding these disclosures, in part because many descriptions are not specific to the model (e.g., central or local) being used. In contrast to this work, we build new explanations rather than evaluating existing ones. We draw on their framework to present the implications of DP (entities to whom data can potentially be disclosed) as part of our nutrition labels. For improved clarity, in our study, we combine two of the disclosures (organization and data analyst), resulting in five total (Table 1).

Table 1: Five Information Disclosures. We combine the "organization" and "data analyst" categories from prior work, since a data analyst is simply an employee of the organization [18]. Although some implementations of the central model limit employees' access to the data (e.g., Uber [47]), we consider the more common case where only published information is privacy-protected.

Information Disclosure	Local	Central
Hack: *A criminal or foreign government that hacks the non-profit could learn my medical history.*	False	True
Law: *A law enforcement organization could access my medical history with a court order requesting this data from the non-profit.*	False	True
Org: *An employee working for the non-profit, such as a data analyst, could be able to see my exact medical history.*	False	True
Graph: *Graphs or informational charts created using information given to the non-profit could reveal my medical history.*	False	False
Share: *Data that the non-profit shares with other organizations doing medical research could reveal my medical history.*	False	True

While Xiong et al. [106] developed explanations that highlight two of these disclosures—namely, disclosure to hackers and disclosure through aggregate statistics (i.e., graphs and charts)—the other information disclosures in Table 1 are not discussed. While some of these information disclosures are equivalent from a technical perspective, prior work suggests that these connections are not clear to a lay audience [18]. In other words, lay audiences may need these implications stated more explicitly. Therefore, unlike Xiong et al. [106], we communicate a wider range of information disclosures. We also differ from Xiong et al. [106] in our approach to evaluation: while they use willingness to share as a measure of explanation quality, we adopt a range of evaluation measures with a focus on comprehension of implications. We take the perspective that a good explanation does not necessarily always increase willingness to share.

Finally, Frazen et al. [34] and Nanayakkara et al. [76] developed methods of explaining the implications of the privacy budget, drawing from the risk communication literature. Nanayakkara et al. [76] found that participants were more willing to share information as the privacy loss budget decreased (i.e., protections were strengthened). In our study, we assume a small privacy loss budget (i.e., strong privacy protections), so that we can focus on the implications of the deployment model. Future work could consider combining our explanations with explanations of the privacy loss budget.

Metaphors. Metaphors are one approach for improving mental models and have been studied extensively in the S&P domain [21, 50, 82, 94, 109, 110]. For example, physical security metaphors can improve users' understanding of personal firewalls [82]. In other cases, however, metaphors have been less effective. For example, descriptions of end-to-end encryption using metaphors failed to improve understanding [21]. Prior work has also begun to explore the effectiveness of metaphors specifically for explaining DP [50]. They find that functional metaphors can be useful for explaining both that injected randomness protects privacy and that there exists a tradeoff between privacy and accuracy. The metaphors we develop are also functional (i.e., focused on *what* DP offers), rather than structural (i.e. focused on *how* DP works) [3]. While the metaphors from prior work aim to cover a long list of facts about DP, they are not designed to emphasize the different kinds of disclosures against which DP may or may not protect—the focus of our work.

Diagrams. Another strategy for explaining PETs is the use of visualizations. For example, hypothetical outcome plots [44] have been used to visualize the protection offered by DP [81, 91]; they have also been used to visualize DP's accuracy implications for data curators [75]. In the case of randomized response [102]—a simple instantiation of local DP—the injected noise can be represented through a spinner [12, 20]. Recent work has also explored the use of diagrams and animations in the specific context of location privacy [107]. Diagrams have also been used to explain other PETs such as encryption [93]. We build on this prior work to develop diagrams for DP in both the local and central models.

Nutrition Labels. One influential approach in privacy communication broadly has been the use of "nutrition labels" for privacy [52]. Drawing inspiration from standardized nutrition labels on food products, privacy labels have been proposed as an alternative or supplement to typical privacy policies with their notorious usability issues [80, 98]. Privacy labels have proven to be a useful way to present privacy-related information [52]. Organizing key information into carefully-designed labels helps users find information more quickly than they would by perusing a traditional privacy policy [53]. Although originally proposed for websites, similar labels have since been developed for datasets [43] and Internet of Things devices [29]. Nutrition labels have even been proposed for describing DP [18, 105]. Apple has recently integrated the nutrition labels approach into their iOS app ecosystem. Unfortunately, the utility of these labels has been hampered by the fact that labels are not always easy to find and can be misleading or inaccurate [17, 57]. Our work adapts the privacy label concept for the purpose of explaining DP—specifically for explaining how local and central DP may or may not protect against particular disclosure risks.

Mental Models. Mental models refer to simplified versions of complex processes that people mentally hold and which may help them understand key pieces of information [16, 62]. Researchers have argued that mental models are important for effectively communicating security risks to end users [92]. Camp [13] argues that a medical or public health mental model is particularly useful for conveying the implications of malicious code—in particular, "that everyone is at risk," "the importance and continued autonomy in the face of risk" and the "shared responsibility for community health." In this way, mental models can rely on people's existing knowledge to help them better grasp attributes of a new setting.

However, flawed mental models can lead to dangerous decisions [100, 103]. For example, Wash [103] proposes folk models of viruses and hackers and describes how these models help explain why people ignore security advice. A mental models approach can also clarify how people's backgrounds may influence their understanding of risks [11, 49, 79]. For instance, people's level of computer science background affects the complexity of their internet mental models, and therefore the number of privacy threats

they perceive [49]. Oates et al. [79] find that when asked to create illustrations of the meaning of privacy, experts' illustrations tend to depict privacy as more "nuanced" than non-experts' illustrations. Bravo-Lillo et al. [11] also find that novice and advanced users have different mental models and risk perceptions.

Finally, researchers have noted the value of studying privacy expectations [62, 83]. For example, Lin et al. [62] propose evaluating mobile app privacy by studying people's privacy expectations of apps, while Rao et al. [83] suggest that understanding misalignments between people's expectations and privacy policies can help reduce privacy risks.

3 Interview Study

We began designing explanations by developing a set of initial prototypes, drawing from prior work in S&P communication. Through an interview study,[2] we use these prototypes to solicit feedback on what makes an effective explanation of DP.

Scenario. We situate our designs within the medical data collection scenario from [18]. In this scenario, a non-profit organization is collecting health data for medical research. Because medical information is considered highly sensitive [45, 89], data subjects are more likely to care about understanding the privacy implications of DP. At the same time, this sensitive data may have the potential to save lives when shared with researchers. The medical scenario provides a particularly powerful example of the tension between privacy and utility—the tension that DP was designed to address.

3.1 Initial Prototypes

Metaphors can help non-experts develop more useful mental models. A candidate list of metaphors was generated by the research team, and the list was iteratively reduced in scope through team discussions. We settled on four initial metaphors—two for the local model and two for the central model—designed to clarify the kinds of risk involved. The candidate metaphors give hints as to how DP works. For example, the idea of blurring an image is similar to adding noise to collected data. However, the focus remains on clearly communicating the implications—for example, that someone with access to the collected data would only see blurry images (in the case of the local model). Thus, while the metaphors were developed with DP in mind, due to the focus on implications, it may be possible to adapt them for other PETs with similar goals. All four metaphors can be found in Appendix D.

We also draw inspiration from prior work on visualizations of DP [12, 24, 75–77, 81, 91] to design our own diagrams that highlight how DP protects or fails to protect against the disclosures listed in Table 1. We developed our diagrams through an iterative process. We discussed the accuracy and clarity of initial diagrams as a group, and based on the discussion, iterated on our designs. In the end, we developed four diagrams—two for the local model and two for the central model—with slight differences in iconography. After initial interviews, we added a third variation for both models that included a caption. All diagrams used a vertical line to depict the "privacy barrier," as in [77], and used icons—most selected from the Noun Project[3]—to represent the different kinds of disclosures. Instead of using an illustration of a database, as in [77, 107], we use an icon of a filing cabinet to represent the collected data. Representative diagrams can be found in Appendix D. While the wavy line passing through the labeled privacy barrier in our diagrams is meant to represent the process of adding noise to data, the diagrams were designed to call attention to information flows rather than to explain how DP works. Thus, again, the focus is on communicating implications of DP.

Following guidance from prior work, we also developed privacy labels to clearly demonstrate which kinds of information disclosures DP can protect against. Each row corresponds to a specific information disclosure and clarifies whether protection is offered against said disclosure. We tested three different versions of the tables (six distinct tables in total, across the two models). One version of the table listed only the disclosures against which DP can protect. Thus for the local model, this table had five rows, whereas for the central model, this version had only one row. This table uses a red circle-backslash symbol to indicate that a particular disclosure is not permitted. The other two versions always included information about all five disclosures, but used different iconography to depict protection or lack thereof. Both of these versions incorporated lock icons to indicate when DP protected against a particular kind of disclosure. In one of these tables, we use a green lock icon—as recommended in prior work on connection security icons [32]—to indicate safety, whereas a red unlocked icon indicates disclosures against which DP does not protect. The other table is in black-and-white and uses the presence or absent of a lock icon to indicate (lack of) protection. Chrome previously used lock icons to indicate connection security, but has recently backed away from this choice due to concerns about overtrust; some Chrome users incorrectly assumed that a lock icon was a reflection on the safety of the website itself rather than the connection [15, 65]. Varying the use of icons allowed us to evaluate their appropriateness in a DP context. Appendix D includes representative versions of our original privacy labels.

3.2 Protocol

We used a 3 x 2 study design: each participant evaluated either the metaphors, diagrams, or privacy labels for either the local or central model. Our goal was to solicit feedback to help us iterate on our designs of each type. All interviews began by describing the same hypothetical scenario:

> *A non-profit organization is asking patients around the country to share their medical records, which will be used to help medical research on improving treatment options and patient care. The non-profit would like to explain to people how they will protect patients' privacy.*

Next, participants are informed that the non-profit plans to "use an extra layer of privacy protection in order to protect patients' medical information." Then, they are shown the first explanation of this privacy protection. After reading the explanation, the participants are asked to explain how patient data will be protected in their own words, as in [37]. Next, they are asked how they feel about the explanation, how well they feel that they understand the privacy protection after reading the explanation, what concerns

[2] All study protocols were reviewed by the authors' institutions' Human Research Protection Office and were determined to be exempt from full IRB review.

[3] https://thenounproject.com

they would have about sharing their data, and what else they would like to know about how patient data will be protected, adapting questions from [84]. If the design under discussion includes the use of color, they are also asked about these color choices. Finally, they are asked how the explanation could be improved.

Next, participants are shown an alternate version of the explanation of the same type, still describing the same model (i.e., local or central). They are asked if the new explanation has changed their understanding. Then, they are asked the same questions they were asked about the original explanation. Some participants were then shown a third version—since we had three versions of the privacy labels and added a third version of the diagrams—and the above questions were repeated. We vary the order of explanations shown between participants. After viewing all versions, participants are asked which one would be most useful for patients deciding whether to share their data. Finally, participants are asked how they would explain to patients how their data would be protected.

Participants who viewed the privacy label or metaphor explanations were then asked to draw a diagram that conveyed their understanding of how patient data would be protected. Participants who struggled to draw on their screens could choose to tell the interviewer what to draw. The purpose of these drawings is two-fold. The drawings serve both as a way to clarify participants' mental models and as a source of inspiration for iterating on our own designs. The participants who viewed the diagram explanations were not asked to do any drawing, since they would be heavily biased towards the diagrams they had already been shown. Finally, in concluding the interview, participants were asked to self-report gender, race, and ethnicity. Additional demographic information was provided through the recruitment platform.

3.3 Participant Recruitment

The first author interviewed 24 U.S. residents (four per condition) recruited through Prolific. Of these 24 participants, eight reviewed metaphors, eight reviewed diagrams, and eight reviewed privacy labels. In each group, half of participants reviewed designs for the local model and half for the central model. All Prolific users are required to be at least 18 years of age. We wanted our explanations to be broadly accessible, so we used Prolific's demographic filters to ensure that at least half of participants had no college degree. A breakdown of participant demographics can be found in Appendix C. Participants whose interviews included drawing a diagram were paid $15, whereas participants who evaluated the diagrams were paid $12 since these interviews were shorter. Interviews lasted about 10-30 minutes and were conducted over Zoom.

3.4 Analysis

Interviews were transcribed using Trint's[4] automated transcription software. The interviewer manually corrected these transcriptions as necessary and created interview summaries, which were discussed with the full research team to validate when saturation was reached sufficiently to proceed with the generation of a codebook and full coding; a final determination that saturation had been reached was made during the coding process (see below). We followed a "collaborative live coding" process [74] in which two or more researchers met together on Zoom to complete each step of the qualitative coding process: from codebook creation through coding of all interviews. First, the lead researcher selected an interview from each condition at random to form a set of six interviews to use to create the codebook. Second, the first two authors met to review these six transcripts and used an inductive approach [10] to develop a set of codes and an initial organization of those codes into themes. Third, these two authors met with the full research team to review and debate the codes, with the result that several codes were revised and additional themes were introduced to better organize the codes based on the group's feedback; this revised codebook organized codes into four distinct themes (Appendix A). Fourth, the same two authors engaged in a series of collaborative zoom meetings during which they coded all 24 interviews together. This collaborative process allowed real-time discussion of disagreements as well as retrospective assessment [38] of whether data saturation had been reached after the coding process was complete.[5]

3.5 Findings

3.5.1 Effectiveness of Initial Designs. While we found some strategies more effective than others, across all conditions, participants had additional questions that our explanations did not answer.

Metaphors. Responses to the metaphors were mixed. Some participants appreciated the concision of the metaphors, while others wanted more details. For example, one participant criticized an explanation's brevity, saying it is *"a little bit simple and [...] doesn't go into too many details." (P8)* In contrast, a different participant complimented this very quality by describing an explanation as *"reader-friendly, very concise"* (P5). This tension between accuracy and thoroughness of explanations on the one hand, and simplicity on the other has also been reported in other domains, such as explainable machine learning [1] and privacy policies [36].

Explanations that make use of metaphor can help people develop useful mental models, and, conversely, people's use of metaphor can reveal their own understanding. Participants across all conditions provided a range of metaphors conveying their understanding of DP, some of which could be adapted as explanations of DP. For example, a participant in the metaphor condition explained that after their data passed through the privacy barrier, they would be like a ghost, no longer identifiable. Another participant in the metaphor condition explained the obfuscation applied in the local model as follows:

> *I have long hair, but you don't know what color it is. You don't know that I have contacts and not glasses, so you wouldn't be able to pick me out of a lineup, is what I would imagine it as.* (P4)

These metaphors of ghosts and lineups both hint at the idea of DP as a form of anonymization. This same participant provided another particularly creative metaphor:

> *It's kind of like an egg. You know, you crack it open and you don't know if it's going to be rotten inside or not. But I don't know what chicken it came from, so I can't blame the chicken.* (P4)

[4] https://trint.com

[5] We do not calculate or report inter-rater reliability (IRR) for two reasons. One, while calculating IRR can be useful to establish agreement before researchers divide a corpus to code different subsets individually, in our case both researchers coded all of the data together. Two, we are not seeking to make quantitative claims about our codes [68].

The phrase "can't blame the chicken" seems to convey the protection offered by DP as a form of plausible deniability.

> **Design Changes:** We replaced our original metaphors with a new metaphor inspired by those generated by participants. Synthesizing metaphors related to hiding or changing one's appearance—like not being recognizable in a lineup or becoming a ghost—we developed a new metaphor: this metaphor compares protecting data with DP to wearing a "disguise."

Diagrams. Of all the explanation methods, the diagrams were the least successful. Of the eight participants assigned to this condition, five explicitly expressed that the diagram was confusing. Although the other three participants did not explicitly use the term "confusing," they also struggled to understand various aspects of the diagrams. For example, when asked to explain the privacy protection in their own words, one participant started to try to explain, then cut themselves off and responded: *"Well, I don't really know"* (P23).

A number of participants expressed confusion or disagreement with the underlying threat model, particularly for the central model. The central model only prevents disclosure from published reports. Although responsible data collectors will employ other technologies such as encryption to protect against hackers or criminals, DP in itself does not protect against this kind of disclosure in the case of the central model. For some participants, this was counterintuitive. For example, after viewing a diagram explaining the central model, one participant expressed their confusion as follows:

> *I don't really get it. [· · ·] There's supposed to be a barrier between my medical information and the people who read the published reports. It seems. And then people who want your data seems like that's open and free, and it seems backwards to me.* (P24)

Two other participants viewing diagrams for the central model incorrectly stated that the privacy barrier was protecting data from hackers, even though the diagrams showed hackers on the left side of the privacy barrier (i.e., the same side as the data collection). One of these participants realized their mistake later in the interview. First, they explained:

> *The privacy barrier [· · ·] allows the people who utilize the information, say the law enforcement and medical professionals, [· · ·] to share that information amongst themselves on a secure in a secure network without allowing the people who want to get that information to abuse that information, the hackers.* (P22)

However, a bit later, they realized their mistake:

> *I'm looking at it again. It says well the people want that data, it's just letting them take it, it looks like. So I guess that would kind of be a concern there [· · ·] we're letting the scientists and the policymakers, the scientists, the people who need to see maybe medical data not allowing them to see the data. But it has a backdoor that allows the people who want to steal that information. So it really has a flaw.* (P22)

Despite the fact that the diagrams showed the hacker to the left of the privacy barrier, two of the four participants in this condition nevertheless explicitly stated that the privacy barrier would protect their data from hackers. Many people may expect PETs to protect against hackers and criminals, making the protection offered by central DP alone somewhat unintuitive [91]. We dropped the diagram explanations due to the pervasive confusion expressed by participants.

Xiong et al. [107] previously investigated the use of diagrams for explaining location privacy and found less than ideal levels of comprehension, particularly for the local model, though they speculate that data quality issues with Amazon Mechanical Turk may be to blame. Alternatively, it is possible that data flow diagrams inherently overemphasize *processes* at the expense of clearly enumerating *implications*.

> **Design Changes:** The diagram explanations were dropped, due to persistent confusion.

Privacy Labels. Responses to the privacy labels were largely positive, though not universally so. Participants praised the privacy labels for their simplicity and clarity. In addition, several participants appreciated the use of color. For example, one participant explained that: "*Having the colored icons does make it a bit faster for a person to get the message*" (P16). However, participants did not always agree about the meaning of the colors green and red. On the one hand, green is often associated with safety while red is associated with danger. Given these associations, one might use green to indicate protection and red to indicate vulnerability. On the other hand, green is also used to mean "go" whereas red means "stop." Given these associations, one might use red to indicate protection, since the flow of data is "stopped." Some participants felt that our use of green and red should be switched, while others felt that our use was appropriate.

> **Design Changes:** To ameliorate the confusion with red and green, we eliminated red and chose to highlight protection in green. The rest of the content was black.

Importance of Process. All our explanations were designed to communicate the *implications* of DP rather than the details of *how* DP works. Prior work finds that explaining the process of adding noise to data is not enough to help people understand the consequences data sharing [106]. Nevertheless, omitting any discussion of process seems to leave people unsatisfied and confused. Most participants had questions about how the data protection worked. Providing a detailed mathematical explanation of DP is likely to overwhelm most people, but people nevertheless do want some information about how DP works—finding the right balance may be challenging. This finding aligns with prior work on explaining encryption. While explanations of encryption focused on *outcome* lead to greater perceived security than explanations focused on *process*, hybrid explanations that incorporate information on both process and outcome lead to the greatest perceived security [26]. Prior work on metaphors for DP also found that some participants were interested in understanding how DP works [50].

> **Design Changes:** Another text was added to provide context about how DP works; we adapted a state-of-the-art text explanation by Xiong et al. [106], with minor adjustments (e.g., rephrasing terms like "database" and "aggregated"). We anticipated that this additional information on *process* could complement our other explanations that focus on *implications*.

3.5.2 Mental Models. Our interviews reveal a number of different mental models that participants constructed to understand DP, based on the explanations they were shown. In many cases, participants' mental models were informed by their prior knowledge of and experience with other technologies.

Comparison to other PETs. Some participants—especially in the diagram and privacy label conditions—reasoned about DP through comparisons to other PETs. For example, one participant understood the privacy barrier as *"some kind of firewall that keeps [their] privacy safe"* (P21). Encryption in particular was mentioned frequently, perhaps because it is a particularly familiar PET or perhaps because participants associated our lock icons with encryption [32, 42]. One participant, who assumed that encryption was the technology being described, wanted to know *"what type of encryption"* (P23) was used. Prior work has also found associations between DP and encryption and has found that associations with encryption correspond to higher trust [50]. DP is distinct from encryption, so while it may be possible to leverage people's knowledge about encryption to construct better explanations of DP, associations with encryption may lead to misconceptions.

One source of confusion is that with encryption, the protection offered should be binary—information is either encrypted (i.e., protected) or not. This corresponds nicely with the physical metaphor of a lock that has exactly two states: locked and unlocked. In the case of DP, however, the goal is to allow some information "leakage" while still offering some protection—the amount of leakage depends on the the privacy budget parameter. Although many participants liked the lock icons, other participants pointed out this issue. For example, one participant in the diagram condition said:

> *If you're releasing some form of my information to these published reports, it's not completely locked.* (P20)

Thus, the use of lock icons and their association with encryption may in some cases prove problematic.

> **Design Changes:** We designed an additional version of the privacy labels that uses arrows to indicate whether data flows are permitted or blocked instead of locks. This version uses red to denote flows that are blocked.

DP as anonymization. Several participants understood DP as an anonymization technique—especially those who read the metaphor explanations. These participants often had an overly-simplistic view of DP. For example, one participant explained that in their understanding, the data *"would be protected by virtue of being anonymized and not including the patient's name, social security number, or date of birth"* (P13). Of course, DP provides better guarantees than such a naive anonymization strategy; nevertheless, this mental model may provide a useful approximation of practical DP guarantees.

DP as fake data. A few participants understood DP as the injection of fake data. One participant explained it as follows:

> *You're collecting my name, but it's a fake one, so it's like a shield up in front of me.* (P4)

Once again, while this model oversimplifies DP, it shares key elements with the truth and thus is likely useful overall. However, it is important for people to understand that the "fake" data nonetheless reveal useful information about the overall distribution; DP does not necessarily protect against inferential privacy risks [54, 55].

3.5.3 Validating Design Changes. We recruited 10 additional participants through Prolific to pilot our updated explanations. These participants were shown explanations of various types—including the two privacy labels and various texts that evolved somewhat over the course of the interviews—and asked to build their own explanation by editing or combining existing explanations or creating their own from scratch (Figure 7). Participants expressed more satisfaction and few substantive edits as compared to our initial evaluations, however they suggested a wide range of ways to combine the texts and privacy labels. No singular combination was preferred by several participants. Therefore, in our quantitative evaluation, we test not only the texts and privacy labels alone but also these explanations in combination with each other as further detailed in Section 4.

Further, prior to launching the full quantitative evaluation of our designs, we compared our two privacy labels in a survey using the evaluative criteria outlined in Section 4. We found no significant differences between the two versions on any of the evaluation criteria. We chose to continue with the version with arrows instead of the version with locks for a few reasons. One participant expressed their preference for the version with arrows as follows:

> *I felt better seeing the same people being blocked rather than the lock because you see those everywhere nowadays.* (P28)

In other words, the lock symbol has become so ubiquitous that this participant found it meaningless. We also felt that the arrows more clearly showed that certain information disclosures were protected against while others were not, whereas lock icons might suggest that certain people are given a "key." This is a fundamentally different kind of protection since keys can be leaked or shared. Finally, although the difference was not significant (analysis details in Section 4), comprehension scores were slightly better for the version with arrows. Thus, we dropped the version with locks. The evolution of our designs is visualized in Figure 8 in Appendix D.

> **Design Changes:** We dropped the label with locks in favor of the version that emphasized information flows.

3.6 Methodological Reflections

Interviews allowed us to solicit detailed feedback from participants, to engage in back and forth discussion, and to understand the nature and source of each misconception. This helped us understand how participants interpreted and felt about our designs so that we could refine them before beginning a large-scale quantitative evaluation. Our insights about participants' mental models may also serve as inspiration for future work on explaining PETs.

A challenge we faced in our interview study was how to deal with conflicting feedback and the fact that certain questions lacked clear consensus. For example, some participants felt strongly that green was the appropriate color for highlighting protections because they associated the color green with safety, whereas other participants felt that red was most appropriate, since the color red indicates that a data flow has been "stopped." While we concluded that it would be best to use only one color so as to minimize the confusion between red and green, we did not have a clear answer as to which color would be the better choice. Similarly, participants did not always agree on what qualifies as a good explanation. For example, some participants requested a great deal of additional information about

both the information collection context and about DP specifically. At the same time, other participants valued simplicity.

4 Large-Scale Evaluation

As a complement to our qualitative evaluation, we conducted an online survey to evaluate our explanations and to assess their efficacy in setting appropriate privacy expectations.

Protocol. Respondents are first asked to read the scenario description (the same medical scenario discussed in Section 3.2) and the description of how data will be protected. Next, respondents answer a simple, multiple-choice comprehension question to ensure that they have read the scenario description. They are given the option to re-read the description. If they do not answer correctly, they are given a second attempt, in accordance with Prolific's policies. If after the second attempt, they again answer incorrectly, they are prevented from advancing further in the survey.

Respondents who pass the comprehension check are then asked whether they trust the non-profit to protect patient privacy [106], followed by two questions related to self-efficacy. Finally, they are asked whether they would be willing to share their information with the non-profit. An open-text box asks them to explain their decision.

Respondents then answer five true/false questions on privacy expectations, followed by the Likert-scale questions about understanding and thoroughness. They are also invited to share feedback on the explanations in a free-response text box. When answering the above questions, respondents have the option to reread the descriptions of the scenario and privacy protection at any time. Next, respondents are asked about their familiarity with various PETs, including DP and a non-existent technology ("deliquescent security"). If they indicate familiarity with some of the listed technologies, they are asked which of the technologies (if any) was described in the survey. In a free-response text box, they are asked to explain their reasoning. Finally, respondents answer questions about themselves. In addition to standard demographic questions (i.e., age, income, race, ethnicity, gender, education, job field), the survey also includes questions to measure internet skill [40]. The internet skill question asks respondents to rate their familiarity with several digital technology concepts (e.g., cache) using a 5-pt scale. The full survey instrument is included in Appendix B.

Experimental Conditions. We use an 8 x 2 experimental design (all conditions listed in Table 2). The levels for explanation type are a control explanation from prior work plus seven new explanation types: privacy label, process text, metaphor text, metaphor + process text, metaphor + process text + privacy label, metaphor text + privacy label, and process text + privacy label. For the control, we use the implications-focused explanations[6] from Xiong et al. [106]. Although [106] evaluated a number of different explanations, we chose to compare specifically against their explanation that led to the highest comprehension of privacy protections. Each respondent was randomly assigned to one condition.

Dependent Measures. Our goal is to set privacy expectations appropriately. Thus, we use a series of true/false questions from prior work about whether certain types of disclosure are possible to measure *objective comprehension* (Table 1). We also ask respondents about their *subjective understanding* of the explanations and how *thorough* they perceive the explanations to be [52]. Additionally, we ask whether respondents *trust* the non-profit organization to "*protect [their] personal information privacy*" [106], and we ask two questions related to *self-efficacy* in decision making [76]. The questions on trust and thoroughness use 5-pt Likert scales. The other three questions use 5-pt semantic scales. Finally, although we do ask about *willingness to share* data with the non-profit as a yes/no question, we caution against using this as a measure of explanation quality. The explanation that convinces the most people to share their data is not necessarily the best explanation. For example, we hope that a patient who is particularly concerned about disclosure to law enforcement would choose *not* to share data when it is protected using the central model. All dependent variables are summarized in Table 3.

Participant Recruitment. 698 total respondents were recruited through Prolific, using the "balanced sample" feature—in accordance with best practices—to recruit an approximately representative sample in terms of gender [95]. We conducted a power analysis to estimate an appropriate sample size; due to the large number of experimental conditions, we lack the statistical power to detect very small effects, but such effects are unlikely to be meaningful in real-world contexts [85]. Respondents were paid $2 for completing the survey, and the median completion time was just under six minutes. Appendix C contains a breakdown of respondent demographics.

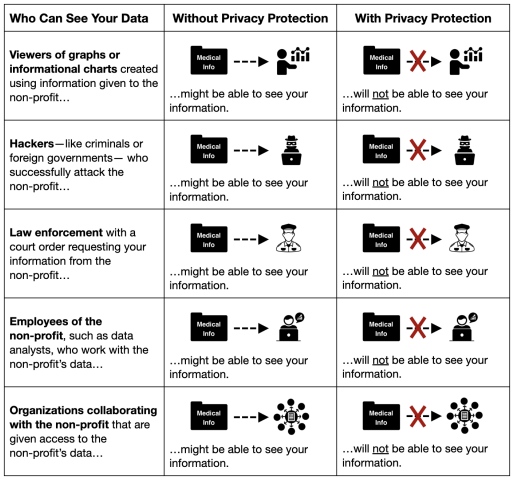
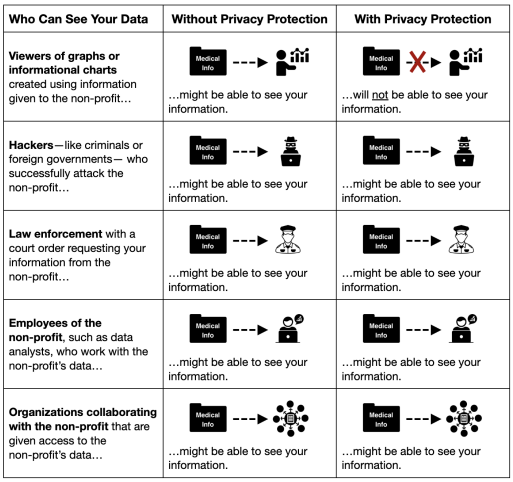
Analysis. We analyze[7] the effect of our explanations on our dependent measures. We construct a set of regression models studying the effect of our independent variables—explanation and model—on our dependent variables: objective comprehension, subjective understanding, perceived thoroughness, trust, self-efficacy, and data-sharing decision. We use logistic regression to study data-sharing decisions, linear regression to study comprehension, and ordinal regression to study the remaining dependent variables. In all models, we control for internet skill [41]. To obtain our internet skill measure, we average each respondent's ratings for the internet skill question in our survey.

We also perform a qualitative analysis of the responses to two of the free-response questions. The first author reviewed all the reasons respondents gave for their data-sharing decisions and developed a set of codes, again employing an inductive approach [10]. The first and second authors then reviewed the codebook together and coded 30 responses, resolving disagreements through discussion and refining the codebook as necessary. Then they separately coded 25 responses and evaluated inter-rater agreement by calculating Cohen's Kappa—the average across all codes appearing in this sample was 0.75, indicating substantial agreement. Remaining responses were divided between both authors for coding. After finding the privacy labels to be most effective, the first author additionally reviewed all feedback provided for the privacy labels and developed a second set of codes—despite some overlap in the themes discussed in the feedback responses and the data-sharing decision responses, the content was sufficiently distinct to merit separate codebooks. Again, the first two authors reviewed the codebook and coded 10 responses together, resolving disagreements through discussion. Then they separately coded 25 responses and

[6] We change the term "app" to "organization" to fit our scenario.

[7] Analysis code: https://osf.io/3acvw/?view_only=f12174861ffd4cd0872a54a8e1326a26

Table 2: All Explanation Texts. Appendix D contains larger versions of the figures.

Type	Local	Central
Arrows Label		
Process	To protect your information, your data will be randomly modified before it is sent to the organization. Only the modified version will be stored, so that your exact data is never collected by the organization.	To protect your information, the organization will store your data but only publish reports, graphs, or charts that have been randomly modified. These modifications hide information that is unique to you as an individual.
Metaphor	The technology works something like this: Your data will be disguised before it is stored by the organization. Therefore, anyone who accesses the data collection will only see this disguised version of your data.	The technology works something like this: The collected data will be disguised when any graphs, charts, or reports are published. However, anyone who accesses the organization's data collection will see the undisguised data.
Metaphor+Process	The technology works something like this: Your data will be disguised before it is stored by the organization. Therefore, anyone who accesses the data collection will only see this disguised version of your data. More specifically, your data will be randomly modified before it is sent to the organization. Only the modified version will be stored, so that your exact data is never collected by the organization.	The technology works something like this: The collected data will be disguised when any graphs, charts, or reports are published. However, anyone who accesses the organization's data collection will see the undisguised data. More specifically, the organization will store your data but only publish reports, graphs, or charts that have been randomly modified. These modifications hide information that is unique to you as an individual.
Label+Metaphor	See *Arrows Label* and *Metaphor* rows.	
Label+Process	See *Arrows Label* and *Process* rows.	
Label+Process+Metaphor	See *Arrows Label* and *Metaphor+Process* rows.	
Xiong et al.	To respect your personal information privacy and ensure best user experience, the data shared with the non-profit organization will be processed via an additional privacy technique. That is, your data will be randomly modified before it is sent to the organization. Since the organization stores only the modified version of your personal information, your privacy is protected even if the organization's database is compromised.	To respect your personal information privacy and ensure best user experience, the data shared with the non-profit organization will be processed via an additional privacy technique. That is, the organization will store your data but only publish the aggregated statistics with modification so that your personal information cannot be learned. However, your personal information may be leaked if the organization's database is compromised.

calculated Cohen's Kappa, with an average of 0.98 across all codes appearing in the sample, indicating near perfect agreement. The remaining 393 responses from participants in any of the privacy label conditions were divided between both authors for coding. For both sets, multiple codes could be applied to a single response. Both sets of codes are available in Appendix A.

4.1 Results

4.1.1 Effectiveness of Designs. We find some explanations are more effective than others in terms of objective comprehension, subjective understanding, and trust.

Comprehension. Across all explanations, we find a significant difference in objective comprehension (Figure 1) between the local and central model—the local model is associated with fewer correct answers ($\beta = -1.17$; $p < 0.01$). This finding is consistent with prior work which suggests that privacy expectations are more closely aligned with the central model than with the local model [18, 107]. It may be difficult to realign a reader's understanding if they come in with strong expectations that do not match the actual protection offered by DP. Compared to the Xiong et al. explanation, all of the explanations that include a privacy label are associated with more correct answers ($\beta = 0.95$–1.26; $p < 0.01$). The text-only explanations, on the other hand, showed no significant improvement over the Xiong et al. explanation. This improvement is expected since the privacy labels are designed explicitly to highlight the information flows that respondents are asked about in the comprehension questions.

Dependent Variable	Survey Item(s)	Adapted From	Model
Objective Comprehension *(Numeric)*	Number of correctly answered true/false questions (Appendix B) about which types of disclosure are possible.	Cummings et al. [18]	linear regression
Subjective Understanding *(Ordinal)*	How confident are you in your understanding of the privacy protection?	Kelley et al. [52]	ordinal regression
Thoroughness *(Ordinal)*	Please indicate your agreement with the following statement: I feel that it was explained thoroughly to me how the non-profit protects patient privacy.	Kelley et al. [52]	ordinal regression
Trust *(Ordinal)*	Please indicate your agreement with the following statement: I trust the non-profit organization to protect my personal information privacy.	Xiong et al. [106]	ordinal regression
Self-Efficacy *(Ordinal)*	How confident are you that you have enough information to decide whether to share your medical record with the non-profit? How confident are you about deciding whether to share your medical record with the non-profit?	Nanayakkara et al. [76]	ordinal regression
Share *(Binary)*	Would you be willing to share your medical record with the non-profit?	Cummings et al. [18]	logistic regression

Table 3: We fit a model for each evaluation measure and for willingness to share: DependentVariable ~ Model + Condition + InternetSkill

.

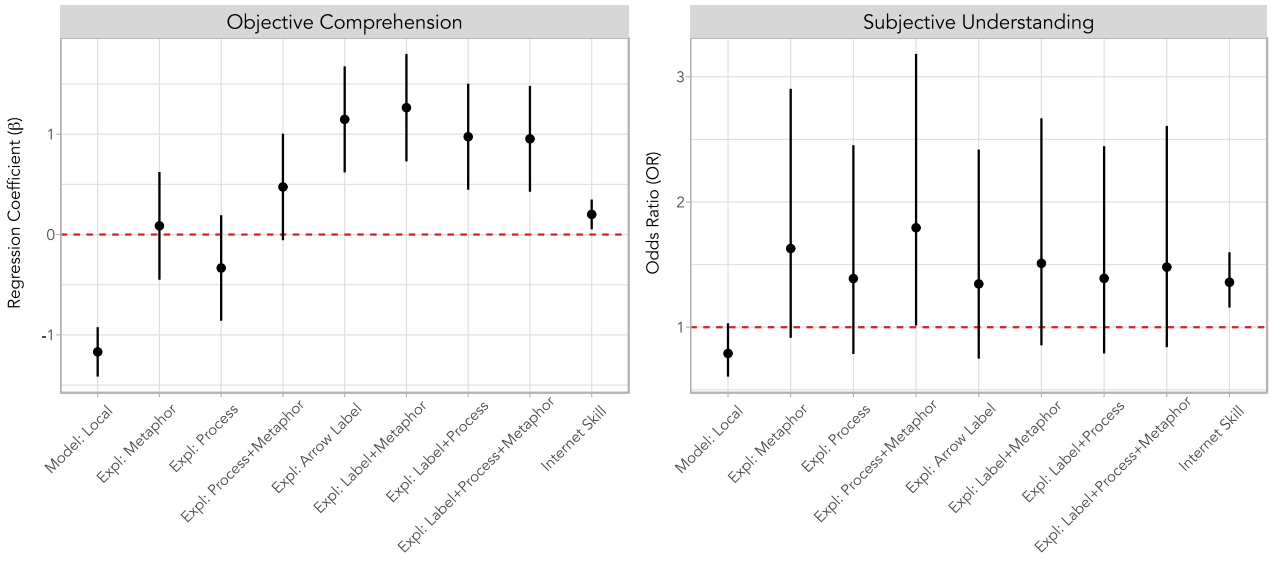


Figure 1: *Left:* Results from linear regression model for objective comprehension. We plot regression coefficients (β) and 95% CIs for these coefficients. $\beta > 0$ indicates an increase while $\beta < 0$ indicates a decrease. *Right:* Results from ordinal regression model for subjective understanding. We plot odds ratios (OR) and corresponding 95% CIs. An OR > 1 indicates an increase in odds, while an OR < 1 indicates a decrease. For both plots, we use the Xiong et al. explanation as the reference level explanation. Table 11 in Appendix E reports the numeric values.

Interestingly, there is a misalignment between objective comprehension and subjective understanding. The process+metaphor explanation is the only one that significantly improves subjective understanding compared to the Xiong et al. baseline (OR = 1.79; $p < 0.05$), even though it does not improve objective comprehension. Prior work has found similar misalignment between objective comprehension and subjective understanding [34, 91]. Unsurprisingly, internet skill is also associated with higher objective comprehension ($\beta = 0.20$; $p < 0.01$) and subjective understanding (OR = 1.36; $p < 0.001$).

Other Evaluation Criteria. Figure 2 summarizes how the explanations compare on our other evaluation criteria. Although comprehension is better for the central model, trust is higher for the local model (OR = 1.97; $p < 0.05$). This is promising, since the local model does offer stronger privacy. The label + process explanation is also associated with greater trust. This aligns with the qualitative feedback from our interviews. While information about process is not enough to help readers understand implications, it seems that explanations that focus only on implications leave readers feeling skeptical. This result is consistent with prior work on explaining encryption that finds benefits of combining information

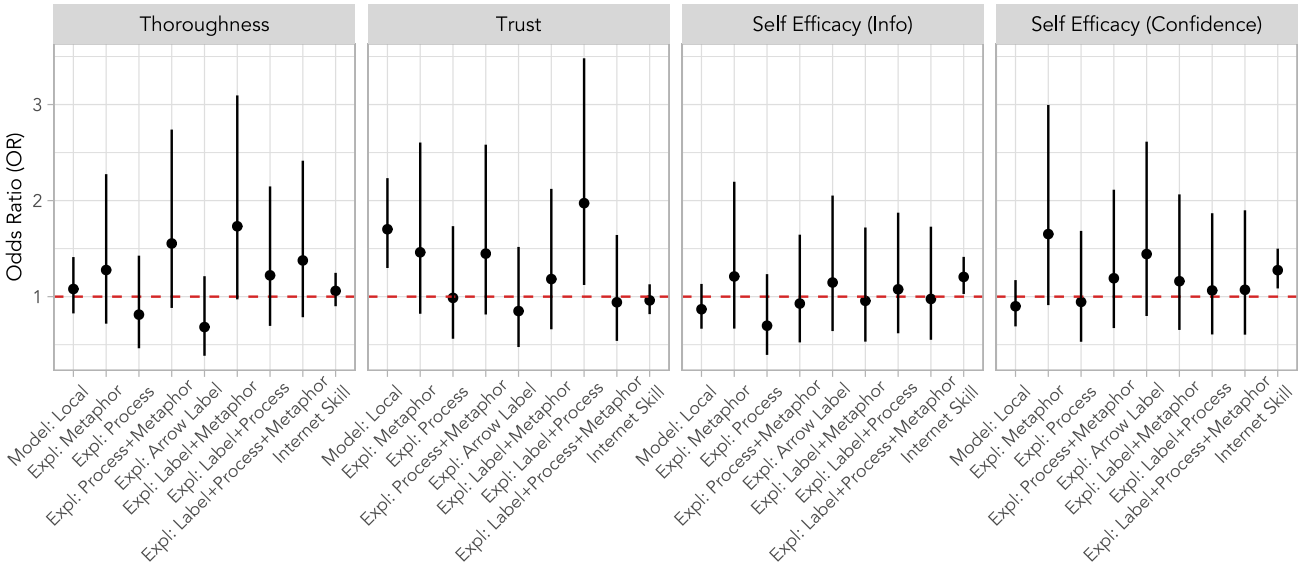


Figure 2: Results from ordinal regression models for trust, perceived thoroughness, and self-efficacy, with the Xiong et al. explanation as the reference level explanation. We plot odds ratios with 95% CIs. An OR > 1 indicates an increase in odds, while an OR < 1 indicates a decrease. Table 12 in Appendix E reports the numeric values.

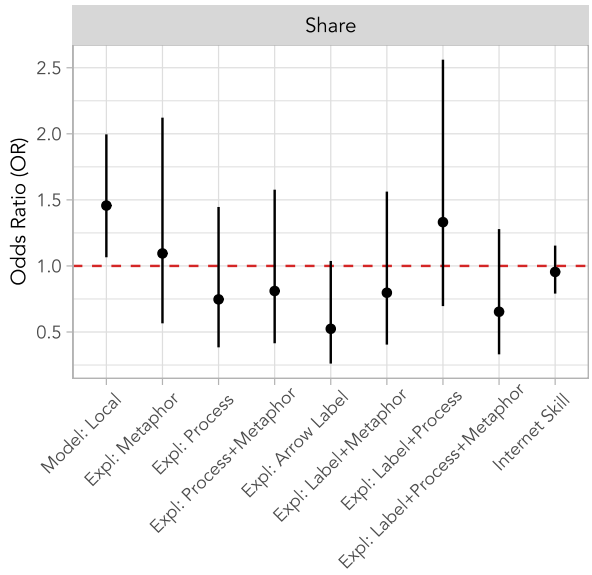


Figure 3: Results from logistic regression model for data-sharing decision. We plot odds ratios (OR) and corresponding 95% CIs. An OR > 1 indicates an increase in odds, while an OR < 1 indicates a decrease. We use the Xiong et al. explanation as the reference level explanation. Table 9 in Appendix E reports the numeric values.

on process and outcome [26]. There were no significant effects of model or explanation on perceived thoroughness or self-efficacy, although higher internet skill is associated with higher self-efficacy (OR = 1.21–1.28; $p < 0.05$).

Feedback. As in our interview study, one of the most common themes in our respondents' feedback was a desire for more information about how the privacy protection works ($n = 76$). For example, one respondent wrote:

> *It doesn't explain at all how this supposed "privacy protection" works, so how do I know if it's credible? I have a lot of cybersecurity training: I want technical details!*

Even respondents who read the process text sometimes requested more information about data protection processes. Respondents also requested other kinds of information ($n = 28$), for example, about the organization and how it would use their data. A tension was again evident between respondents who requested additional information and those who praised our concision or requested further simplification. One respondent suggested *"more detailed explanations of the privacy protections that are available [...] if needed."*

4.1.2 Prior Familiarity with PETs. In interviews, we found that some participants understood DP through comparisons with other PETs. Of the PETs we mention in our survey, end-to-end encryption was by far the most familiar, whereas only a minority had heard of DP (Appendix C). Of respondents who answered the question asking which technology was described in the survey, most correctly selected DP, though several respondents explained in their free-text responses that they were simply guessing.

4.1.3 Data-Sharing Decision. Respondents are more willing to share data (Table 9) under the local model (OR = 1.46; $p < 0.05$).

This replicates findings from prior work and is likely due to the stronger privacy guarantees of the local model [106]. None of the explanations had a significant effect on data-sharing decisions.

When people decide whether to share information, they consider many other factors in addition to privacy protections [35, 73, 91]. In fact, many respondents simply were not worried about privacy (n=75). For example, one respondent felt that they had nothing in their medical history that they would "*need to hide or be particularly private about.*" Other respondents were interested in sharing their information to help others, mentioning benefits of data sharing (n=151). In the words of one respondent: "*I do not have a problem with sharing my records if it will help someone.*" On the other hand, respondents who were less willing to share their data often felt it would be too risky or that their medical information was simply too private (n=242). For example, one respondent explained they were "*not comfortable sharing [their] medical records with anyone but [their] doctor.*" Other respondents wanted more information before they would be willing to share their data (n=155). The information they requested was not always related to DP. For example, some respondents wanted to know more about the non-profit organization. Finally, some participants distrusted either the non-profit or the privacy protection (n=88). In the words of one respondent: "*Companies say that your information is secure all the time, but all the time there are security breaches.*" Other respondents also mentioned the frequency of data breaches as a cause for concern (n=35).

4.2 Methodological Reflections

Just as interview participants disagreed about what makes a good explanation, there is no established standard among researchers for evaluating explanations of DP. We sought to integrate various evaluation measures from prior work. Most studies include some measure of comprehension [18, 34, 59, 76, 91, 106], but researchers may disagree about what information is most important for end users. While we focus on communicating privacy implications, in some contexts, it may be important to communicate accuracy implications as well [6, 33]. Researchers may also disagree about the extent to which explanations should explain how DP works. Yet even if researchers were to agree upon a set of comprehension questions, not all misconceptions are equally harmful. For example, depending on context, it may be worse to overestimate the protection DP offers than to underestimate the protection. Although our additional measures of quality—subjective understanding, perceived thoroughness, trust, and self-efficacy—all come from prior work, high scores on these measures only indicate high explanation quality when paired with high comprehension scores. For example, high scores on trust and subjective understanding could be dangerous if participants have severely misunderstood the protection DP offers.

We see some overlap between the feedback collected in interviews and the open-ended feedback collected through our survey. Most notably, in both the interview transcripts and the survey responses, we see a tension between a desire for more information and a desire for simplicity. Our qualitative analysis of open-ended responses is not a replacement for participant interviews since we cannot, for example, probe survey respondents with follow-up questions to identify the sources of misconceptions. Instead, we view the survey and interview data as complementary.

5 Limitations

Our designs are limited in their focus on a single scenario. Although medical applications are often cited as motivation for studies of DP [8, 48], DP has not been widely deployed in medical contexts [19]. Nevertheless, our privacy labels are transportable to other domains. Future work could transfer our designs to other scenarios and test whether our findings still hold. A limitation of our evaluation is that encountering explanations of DP in practice differs significantly from encountering explanations in an online survey. Future work could investigate comprehension when these explanations are encountered in more natural settings. A third limitation is our focus on a U.S. audience. Our privacy labels may be received differently in a different cultural context. Finally, we present the nature of DP's protection as binary, when in fact the level of protection depends on the choice of privacy budget. This simplification may be appropriate for small privacy budgets, but the question of determining an acceptable range for the privacy budget is itself a nontrivial problem.

Another limitation is that the survey respondents recruited through Prolific may not be truly representative of the US population. In particular, one risk is that these respondents may be unusually "tech-savvy." Most respondents did not work in technical fields (see demographics in Appendix C). Nevertheless, prior work has found that Prolific users are, on average, more knowledgeable about security and privacy than the general US population [95]. Thus, there is some risk that we have underestimated how challenging our explanations might be for people with less knowledge on these topics. Furthermore, it is unclear whether our results might generalize to non-US contexts. For example, while prior work shows that the information disclosures we emphasize matter to US participants [18], they may not be the most salient concerns for participants in other countries. Future work could seek to answer these questions by replicating our study in other cultural contexts.

One concern may be that our privacy labels are "teaching to the test," since we design them specifically to highlight the information disclosures that we ask about to measure comprehension. Thus, it is not surprising that comprehension is higher for our privacy labels than for explanations designed with a different emphasis. However, if the purpose of an explanation is to inform readers about which information flows are restricted—i.e., if we are using the "right" test—perhaps teaching to the test is not such a problem. Nevertheless, we incorporate additional evaluation criteria from prior work and find that our privacy labels improve comprehension without sacrificing quality on these other metrics (Tables 11 – 12).

6 Discussion

Our results highlight the value of combining disparate best practices from prior work on explaining other S&P concepts to explain complex PETs such as DP [26, 52]. We find that consequences-focused explanations—specifically, privacy label explanations that highlight information flows—to be a promising approach for promoting accurate understandings of potential data leaks in DP systems. However, to ensure that such explanations are trusted we find that it is necessary to pair such consequences-focused information with a limited amount of high-level information about mechanisms: how DP works to offer particular consequences and protections.

Below we discuss potential pitfalls of privacy labels for DP as well as ways to extend our designs to explain other PETs individually or in combination.

Potential Pitfalls. Although the nutrition label approach shows promise for setting appropriate privacy expectations, it is important to avoid pitfalls from prior deployments of nutrition labels for privacy [17]. For example, iOS privacy labels can be misleading and inaccurate [57], in part because developers struggle to create accurate labels [61]. Similarly, our labels for DP could be misleading if an organization has implemented DP incorrectly [9, 14, 46, 64, 72] or has chosen an inappropriately large privacy budget [27]. Specialized programming platforms, audits, and formal verification approaches can complement our work by ensuring that the communicated privacy guarantees match the implementation [25, 56, 69, 86, 97, 108].

Furthermore, while privacy labels can empower individuals to make decisions that better align with their goals and values, it is also important not to overburden individuals in the same way that traditional privacy policies do [67]. As some of the participants we interviewed highlighted, it can be difficult to strike the right balance between simplicity and comprehensiveness. Our nutrition labels in particular focus on high-level implications that are not necessarily specific to DP; for example, the label for the local model could easily be adapted to explain other PETs within the paradigm of privacy-preserving outsourced computation (e.g., private federated learning). For most data subjects, understanding the *implications* of DP is more important than understanding data protection *processes*, even though our results show that people are often curious about how DP works. A simple nutrition label paired with a short text description explaining data protection processes may strike the right balance. This process description would contain information that is specific to DP and could be expanded as necessary depending on the context.

Such a balance between simplicity and comprehensiveness is important not only for data subjects, but also for other audiences who may encounter DP. For instance, privacy labels for DP could be used to educate policymakers, advocacy organizations, or software developers to support them in various decision-making processes. For example, Mozilla's "privacy not included" guide offers expert reviews to help buyers choose products that provide strong privacy and security, since it can be difficult for individual buyers to evaluate various data protection policies themselves. One could imagine a similar project to provide reviews for different data collection initiatives. An advocacy organization might use privacy labels for PETs like DP to identify and recommend certain initiatives that provide good S&P guarantees.

Finally, it is crucial that privacy labels for DP be contextual. While the information disclosures our explanations highlight are ones that people care about [18], they represent a starting point which should be used to further adapt explanations for specific contexts. The information disclosures we highlight may not be comprehensive of all specific disclosures people are concerned about across contexts. For example, privacy concerns in a particular educational setting may differ from a medical setting. Future work should also study ways to supplement privacy labels for DP with contextually-appropriate communication about the choice of privacy budget [7, 76].

Privacy Labels for Other PETs. Our approach to privacy labels for DP could be adapted to other PETs. We hypothesize that privacy labels that take a contextual integrity approach—emphasizing which data flows are permitted and which are prohibited—could lead to improved comprehension of a variety of PETs [78]. Our survey respondents found it more difficult to reason about the implications of local DP than central DP. This finding suggests that clearly explaining which data flows are permitted is particularly important for PETs that enable outsourced computation, such as local DP. Future work could confirm whether the techniques employed here, and the greater difficulty with mental model formation among participants, extends to other PETs that engage in outsourced computation, such as secure multi-party computation.

Our findings suggest that people employ their known models of PETs (e.g., understandings of encryption) to reason about new PETs. A standardized approach for presenting the kinds of protection a particular PET offers could help people compare new PETs with more familiar ones. Leveraging this kind of prior knowledge could be beneficial; however, we also caution that in some cases, drawing on knowledge of other PETs could lead to confusion or overtrust. It is important that comparisons between PETs clearly explain their differences and do not overstate the protection offered.

Finally, PETs are rarely deployed in isolation. Our qualitative data show that people are interested in learning about DP *in context*. That is, they want information about the protection offered by DP, but they also care about the other safeguards and signals of trustworthiness that might help them make better-informed holistic data-sharing decisions. Particularly in the case of the central model, users may feel more comfortable if information about DP is presented alongside information about other PETs used to secure user data. Future work should go beyond explaining PETs one at a time and study effective ways to explain the nature of the protection obtained through combinations of PETs. Since our privacy labels focus on information flows—rather than the details of how DP works—it should be straightforward to modify them to communicate the protection offered by multiple PETs in combination.

Evaluating Explanations. Finally, we suggest that there is a need for further exploration and discussion on best practices for evaluating explanations of PETs. We have noted some of the tradeoffs between qualitative versus quantitative evaluations. While qualitative data allows for a deeper and more nuanced understanding of how explanations are interpreted, there may be a risk of overfitting to the feedback of a small sample. Prior work suggests viewing quantitative and qualitative methods as complementary approaches [24, 60, 66, 90, 99]. We employ a mixed-methods approach to take advantage of the strengths of both qualitative and quantitative data. Yet the field does not have well-established standards for either qualitative or quantitative evaluations. Additional research and discussion among researchers is needed (1) to determine what qualities make an explanation effective and (2) to develop valid metrics to measure the relevant qualities. The metrics we adopt in this study are informed by prior work and may serve as a starting point for future studies, but additional or modified metrics may be necessary depending on context. For example, the kinds of comprehension questions that are appropriate for data subjects may not be appropriate for other audiences, such as data curators.

Acknowledgments

We would like to thank everyone who provided feedback on various stages of this project, especially Aaron Broukhim and participants in the Technically Private reading group. All authors were supported by DARPA (contract number W911NF-21-1-0371). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Government or DARPA. In addition to DARPA support, the first author was supported in part by NSF Grant #2429838, the third author was supported in part by NSF grant CNS-1942772 (CAREER), a Mozilla Research Grant, a JP-Morgan Chase Faculty Research Award, and an Apple Privacy-Preserving Machine Learning Award. The fourth author was also supported by NSF grant #2030859 to the Computing Research Association for the CIFellows Project, and the fifth author was also supported by a Google Research Scholar Award and NSF grant #2429838.

References

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] John M. Abowd. 2018. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2867. https://doi.org/10.1145/3219819.3226070

[3] Ala Sarah Alaqra, Farzaneh Karegar, and Simone Fischer-Hübner. 2023. Communicating the Privacy Functionality of PETs to eHealth Stakeholders. (2023).

[4] Differential Privacy Team Apple. 2017. Learning with privacy at scale. *Apple Machine Learning Journal* 1, 8 (2017).

[5] Farzaneh Asgharpour, Debin Liu, and L. Jean Camp. 2007. Mental Models of Security Risks. In *Financial Cryptography and Data Security*, Sven Dietrich and Rachna Dhamija (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 367–377.

[6] Narges Ashena, Oana Inel, Badrie L Persaud, and Abraham Bernstein. 2024. Casual users and rational choices within differential privacy. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 932–950.

[7] Sebastian Benthall and Rachel Cummings. 2022. Integrating Differential Privacy and Contextual Integrity. USENIX Association, Santa Clara, CA.

[8] Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. 2010. Discovering frequent patterns in sensitive data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 503–512.

[9] Benjamin Bichsel, Samuel Steffen, Ilija Bogunovic, and Martin Vechev. 2021. DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifiers. In *2021 IEEE Symposium on Security and Privacy (SP)*. 391–409. https://doi.org/10.1109/SP40001.2021.00081

[10] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. Sage.

[11] Christian Bravo-Lillo, Lorrie Faith Cranor, Julie S. Downs, and Saranga Komanduri. 2011. Bridging the Gap in Computer Security Warnings: A Mental Model Approach. *IEEE Security & Privacy* 9, 2 (March 2011), 18–26.

[12] Brooke Bullek, Stephanie Garboski, Darakhshan J. Mir, and Evan M. Peck. 2017. Towards Understanding Differential Privacy: When Do People Trust Randomized Response Technique?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, Denver, Colorado, USA, 3833–3837. https://doi.org/10.1145/3025453.3025698

[13] L Jean Camp. 2009. Mental models of privacy and security. *IEEE Technology and society magazine* 28, 3 (2009), 37–46.

[14] Sílvia Casacuberta, Michael Shoemate, Salil P. Vadhan, and Connor Wagaman. 2022. Widespread Underestimation of Sensitivity in Differentially Private Libraries and How to Fix It. In *ACM CCS 2022*, Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi (Eds.). ACM Press, 471–484. https://doi.org/10.1145/3548606.3560708

[15] Chromium Blog. 2023. An Update on the Lock Icon.

[16] Kenneth James Williams Craik. 1967. *The nature of explanation*. Vol. 445. CUP Archive.

[17] Lorrie Faith Cranor. 2022. Mobile-app privacy nutrition labels missing key ingredients for success. *Commun. ACM* 65, 11 (2022), 26–28.

[18] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2021. "I need a better description": An Investigation Into User Expectations For Differential Privacy. In *ACM CCS 2021*, Giovanni Vigna and Elaine Shi (Eds.). ACM Press, 3037–3052. https://doi.org/10.1145/3460120.3485252

[19] Fida Kamal Dankar and Khaled El Emam. 2013. Practicing differential privacy in health care: A review. *Trans. Data Priv.* 6, 1 (2013), 35–67.

[20] Inbal Dekel, Rachel Cummings, Ori Heffetz, and Katrina Ligett. 2023. The Privacy Elasticity of Behavior: Conceptualization and Application. In *Proceedings of the 24th ACM Conference on Economics and Computation (EC '23)*.

[21] Albese Demjaha, Jonathan M Spring, Ingolf Becker, Simon Parkin, and M Angela Sasse. 2018. Metaphors considered harmful? An exploratory study of the effectiveness of functional metaphors for end-to-end encryption. In *Proc. USEC*, Vol. 2018. Internet Society.

[22] Damien Desfontaines. 2021. A list of real-world uses of differential privacy. https://desfontain.es/privacy/real-world-differential-privacy.html. Ted is writing things (personal blog).

[23] Damien Desfontaines and Balázs Pejó. 2020. Sok: differential privacies. *Proceedings on privacy enhancing technologies* 2020, 2 (2020), 288–313.

[24] Onyinye Dibia, Brad Stenger, Steven Baldasty, Mako Bates, Ivoline C. Ngong, Yuanyuan Feng, and Joseph P. Near. 2024. SoK: Usability Studies in Differential Privacy. arXiv:2412.16825 [cs.HC] https://arxiv.org/abs/2412.16825

[25] Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. 2018. Detecting Violations of Differential Privacy. In *ACM CCS 2018*, David Lie, Mohammad Mannan, Michael Backes, and XiaoFeng Wang (Eds.). ACM Press, 475–489. https://doi.org/10.1145/3243734.3243818

[26] Verena Distler, Carine Lallemand, and Vincent Koenig. 2020. Making encryption feel secure: Investigating how descriptions of encryption impact perceived security. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 220–229.

[27] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. 2019. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality* 9, 2 (2019).

[28] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC 2006 (LNCS, Vol. 3876)*, Shai Halevi and Tal Rabin (Eds.). Springer, Berlin, Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[29] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2022. An Informative Security and Privacy "Nutrition" Label for Internet of Things Devices. *IEEE Security & Privacy* 20, 02 (2022), 31–39.

[30] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 1054–1067.

[31] Matthias Fassl, Lea Theresa Gröber, and Katharina Krombholz. 2021. Exploring User-Centered Security Design for Usable Authentication Ceremonies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 694, 15 pages. https://doi.org/10.1145/3411764.3445164

[32] Adrienne Porter Felt, Robert W Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. 2016. Rethinking connection security indicators. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. 1–14.

[33] Daniel Franzen, Claudia Müller-Birn, and Odette Wegwarth. 2024. Communicating the privacy-utility trade-off: Supporting informed data donation with privacy decision interfaces for differential privacy. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–56.

[34] Daniel Franzen, Saskia Nuñez von Voigt, Peter Sörries, Florian Tschorsch, and Claudia Müller-Birn. 2022. Am I Private and If So, how Many?: Communicating Privacy Guarantees of Differential Privacy with Risk Communication Formats. In *ACM CCS 2022*, Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi (Eds.). ACM Press, 1125–1139. https://doi.org/10.1145/3548606.3560693

[35] Alisa Frik, Julia Bernd, and Serge Egelman. 2023. A model of contextual factors affecting older adults' information-sharing decisions in the US. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–48.

[36] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Symposium on Usable Privacy and Security (SOUPS '16)*. USENIX, Denver, Colorado, USA, 321–340.

[37] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. 2018. " What was that site doing with my Facebook password?" Designing Password-Reuse Notifications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 1549–1566.

[38] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82.

[39] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. 2018. Away from prying eyes: Analyzing usage and understanding of private browsing.

In *Fourteenth symposium on usable privacy and security (SOUPS 2018)*. 159–175.
[40] Eszter Hargittai and Yuli Patrick Hsieh. 2012. Succinct survey measures of web-use skills. *Social Science Computer Review* 30, 1 (2012), 95–107.
[41] Eszter Hargittai and Marina Micheli. 2019. Internet skills and why they matter. *Society and the internet: How networks of information and communication are changing our lives* 109 (2019).
[42] Amir Herzberg and Hemi Leibowitz. 2016. Can Johnny finally encrypt? Evaluating E2E-encryption in popular IM applications. In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*. 17–28.
[43] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1.
[44] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.
[45] Iulia Ion, Niharika Sachdeva, Ponnurangam Kumaraguru, and Srdjan Čapkun. 2011. Home is safer than the cloud! Privacy concerns for consumer cloud storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*. 1–20.
[46] Jiankai Jin, Eleanor McMurtry, Benjamin I. P. Rubinstein, and Olga Ohrimenko. 2022. Are We There Yet? Timing and Floating-Point Attacks on Differential Privacy Systems. In *2022 IEEE Symposium on Security and Privacy (SP)*. 473–488. https://doi.org/10.1109/SP46214.2022.9833672
[47] Noah Johnson, Joseph P Near, Joseph M Hellerstein, and Dawn Song. 2020. Chorus: a programming framework for building scalable differential privacy mechanisms. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 535–551.
[48] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.
[49] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. 2015. "My data just goes Everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*. Ottawa, 39–52.
[50] Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. 2022. Exploring {User-Suitable} Metaphors for Differentially Private Data Analyses. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. 175–193.
[51] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM J. Comput.* 40, 3 (2011), 793–826.
[52] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. 1–12.
[53] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. 1573–1582.
[54] Daniel Kifer and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 193–204.
[55] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)* 39, 1 (2014), 1–36.
[56] Daniel Kifer, Solomon Messing, Aaron Roth, Abhradeep Thakurta, and Danfeng Zhang. 2020. Guidelines for implementing and auditing differentially private systems. *arXiv preprint arXiv:2002.04049* (2020).
[57] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. 2022. Goodbye tracking? Impact of iOS app tracking transparency and privacy labels. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 508–520.
[58] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel Von Zezschwitz. 2019. " If HTTPS Were Secure, I Wouldn't Need 2FA"-End User and Administrator Mental Models of HTTPS. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 246–263.
[59] Patrick Kühtreiber, Viktoriya Pak, and Delphine Reinhardt. 2022. Replication: The Effect of Differential Privacy Communication on German Users' Comprehension and Data Sharing Attitudes. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. 117–134.
[60] Juha Lehikoinen and Ville Koistinen. 2014. In big data we trust? *Interactions* 21, 5 (2014), 38–41.
[61] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. 2022. Understanding challenges for developers to create accurate privacy nutrition labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.
[62] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. 2012. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 501–510.
[63] Heather Richter Lipford, Gordon Hull, Celine Latulipe, Andrew Besmer, and Jason Watson. 2009. Visible flows: Contextual integrity and the design of privacy mechanisms on social network sites. In *2009 International Conference on Computational Science and Engineering*, Vol. 4. IEEE, 985–989.
[64] Min Lyu, Dong Su, and Ninghui Li. 2017. Understanding the Sparse Vector Technique for Differential Privacy. *Proc. VLDB Endow.* 10, 6 (feb 2017), 637–648. https://doi.org/10.14778/3055330.3055331
[65] Zane Ma, Joshua Reynolds, Joseph Dickinson, Kaishen Wang, Taylor Judd, Joseph D Barnes, Joshua Mason, and Michael Bailey. 2019. The impact of secure transport protocols on phishing efficacy. In *12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 19)*.
[66] Doren L Madey. 1982. Some benefits of integrating qualitative and quantitative methods in program evaluation, with illustrations. *Educational evaluation and policy analysis* 4, 2 (1982), 223–236.
[67] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp* 4 (2008), 543.
[68] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (nov 2019), 23 pages. https://doi.org/10.1145/3359174
[69] Frank D. McSherry. 2009. Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data* (Providence, Rhode Island, USA) *(SIGMOD '09)*. Association for Computing Machinery, New York, NY, USA, 19–30. https://doi.org/10.1145/1559845.1559850
[70] Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins. 2020. Facebook Privacy-Protected Full URLs Data Set. https://doi.org/10.7910/DVN/TDOAPG
[71] Gerome Miklau. 2021. How Tumult Labs helped the IRS support educational accountability with differential privacy. https://www.tmlt.io/research/how-tumult-labs-helped-irs-support-educational-accountability-with-differential-privacy
[72] Ilya Mironov. 2012. On significance of the least significant bits for differential privacy. In *ACM CCS 2012*, Ting Yu, George Danezis, and Virgil D. Gligor (Eds.). ACM Press, 650–661. https://doi.org/10.1145/2382196.2382264
[73] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. 2017. Privacy expectations and preferences in an IoT world. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association Santa Clara, 399–412.
[74] Gayathri Naganathan, Sinthu Srikanthan, Abhirami Balachandran, Angel Gladdy, and Vasuki Shanmuganathan. 2022. Collaborative zoom coding—a novel approach to qualitative analysis. *International Journal of Qualitative Methods* 21 (2022), 16094069221075862.
[75] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica R. Hullman, and Jennie Duggan. 2022. Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases. *Proceedings on Privacy Enhancing Technologies* 2022 (2022), 601 – 618.
[76] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. 2023. What are the chances? Explaining the epsilon parameter in differential privacy. In *32nd USENIX Security Symposium (USENIX Security 23)*. 1613–1630.
[77] Joseph Near and David Darais. 2020. Threat Models for Differential Privacy . https://www.nist.gov/blogs/cybersecurity-insights/threat-models-differential-privacy
[78] Helen Nissenbaum. 2009. Privacy in context. In *Privacy in Context*. Stanford University Press.
[79] Maggie Oates, Yama Ahmadullah, Abigail Marsh, Chelse Swoopes, Shikun Zhang, Rebecca Balebako, and Lorrie Faith Cranor. 2018. Turtles, locks, and bathrooms: Understanding mental models of privacy through illustration. *Proceedings on Privacy Enhancing Technologies* 2018, 4 (2018), 5–32.
[80] Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.
[81] Adam Pearce and Ellen Jiang. 2020. How randomized response can help collect sensitive information responsibly. https://pair.withgoogle.com/explorables/anonymization/
[82] Fahimeh Raja, Kirstie Hawkey, Steven Hsu, Kai-Le Clement Wang, and Konstantin Beznosov. 2011. A brick wall, a locked door, and a bandit: a physical security metaphor for firewall warnings. In *Proceedings of the seventh symposium on usable privacy and security*. 1–20.
[83] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. 2016. Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online. In *Symposium on Usable Privacy and Security (SOUPS '16)*. USENIX, Denver, Colorado, USA, 77–96.
[84] Elissa M Redmiles, Everest Liu, and Michelle L Mazurek. 2017. You Want Me To Do What? A Design Study of Two-Factor Authentication Messages.. In *SOUPS*, Vol. 57. 93.

[85] Elissa M. Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L. Mazurek. 2018. Asking for a Friend: Evaluating Response Biases in Security User Studies. In *ACM CCS 2018*, David Lie, Mohammad Mannan, Michael Backes, and XiaoFeng Wang (Eds.). ACM Press, 1238–1255. https://doi.org/10.1145/3243734.3243740

[86] Jason Reed and Benjamin C. Pierce. 2010. Distance Makes the Types Grow Stronger: A Calculus for Differential Privacy. *SIGPLAN Not.* 45, 9 (sep 2010), 157–168. https://doi.org/10.1145/1932681.1863568

[87] Phillip Rogaway. 2015. The moral character of cryptographic work. *Cryptology ePrint Archive* (2015).

[88] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. 2015. A Design Space for Effective Privacy Notices. In *Symposium on Usable Privacy and Security (SOUPS '15)*. USENIX, Ottawa, Canada, 1–17.

[89] Eva-Maria Schomakers, Chantal Lidynia, Dirk Müllmann, and Martina Ziefle. 2019. Internet users' perceptions of information sensitivity–insights from Germany. *International Journal of Information Management* 46 (2019), 142–150.

[90] Imani N Sherman, Jack W Stokes, and Elissa M Redmiles. 2021. Designing media provenance indicators to combat fake media. In *Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses*. 324–339.

[91] Mary Anne Smart, Dhruv Sood, and Kristen Vaccaro. 2022. Understanding Risks of Privacy Theater with Differential Privacy. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 342 (nov 2022), 24 pages. https://doi.org/10.1145/3555762

[92] Geordie Stewart and David Lacey. 2012. Death by a thousand facts: Criticising the technocratic approach to information security awareness. *Information Management & Computer Security* 20, 1 (2012), 29–38.

[93] Christian Stransky, Dominik Wermke, Johanna Schrader, Nicolas Huaman, Yasemin Acar, Anna Lena Fehlhaber, Miranda Wei, Blase Ur, and Sascha Fahl. 2021. On the limited impact of visualizing encryption: Perceptions of E2E messaging security. In *Seventeenth Symposium on Usable Privacy and Security*. 437–454.

[94] Sangho Suh, Sydney Lamorea, Edith Law, and Leah Zhang-Kennedy. 2022. PrivacyToon: Concept-driven Storytelling with Creativity Support for Privacy Concepts. In *Designing Interactive Systems Conference*. 41–57.

[95] Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. Replication: How well do my results generalize now? The external validity of online privacy and security surveys. In *Eighteenth symposium on usable privacy and security (SOUPS 2022)*. 367–385.

[96] Abhradeep Guha Thakurta, Andrew H Vyrros, Umesh S Vaishampayan, Gaurav Kapoor, Julien Freudiger, Vivek Rangarajan Sridhar, and Doug Davidson. 2017. Learning new words. US Patent 9,594,741.

[97] Florian Tramer, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, and Nicholas Carlini. 2022. Debugging differential privacy: A case study for privacy auditing. *arXiv preprint arXiv:2202.12219* (2022).

[98] Joseph Turow, Michael Hennessy, and Nora Draper. 2018. Persistent misperceptions: Americans' misplaced confidence in privacy policies, 2003–2015. *Journal of Broadcasting & Electronic Media* 62, 3 (2018), 461–478.

[99] Dirk Van Der Linden, Matthew Edwards, Irit Hadar, and Anna Zamansky. 2020. Pets without PETs: on pet owners' under-estimation of privacy concerns in pet wearables. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 143–164.

[100] Kami Vaniea, Emilee Rader, and Rick Wash. 2014. Mental models of software updates. *International Communication Association* (2014), 1–39.

[101] Meridel Walkington. 2019. Designing Better Security Warnings. https://blog.mozilla.org/ux/2019/03/designing-better-security-warnings/

[102] Stanley L. Warner. 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *J. Amer. Statist. Assoc.* 60, 309 (March 1965), 63–69. https://doi.org/10.1080/01621459.1965.10480775

[103] Rick Wash. 2010. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*. 1–16.

[104] Alma Whitten and J Doug Tygar. 1999. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0.. In *USENIX security symposium*, Vol. 348. 169–184.

[105] Aiping Xiong. 2020. Effect of Facts Box on Users' Comprehension of Differential Privacy: A Preliminary Study. In *Proceedings of the Human Factors and Ergonomics Society 2020 Annual Meeting*.

[106] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. 2020. Towards effective differential privacy communication for users' data sharing decision and comprehension. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 392–410.

[107] Aiping Xiong, Chuhao Wu, Tianhao Wang, Robert W. Proctor, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2022. Using Illustrations to Communicate Differential Privacy Trust Models: An Investigation of Users' Comprehension, Perception, and Data Sharing Decision. *ArXiv* abs/2202.10014 (2022).

[108] Danfeng Zhang and Daniel Kifer. 2017. LightDP: Towards Automating Differential Privacy Proofs. In *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages* (Paris, France) *(POPL '17)*. Association for Computing Machinery, New York, NY, USA, 888–901. https://doi.org/10.1145/3009837.3009884

[109] Leah Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. 2013. Password advice shouldn't be boring: Visualizing password guessing attacks. In *2013 APWG eCrime Researchers Summit*. 1–11. https://doi.org/10.1109/eCRS.2013.6805770

[110] Leah Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. 2016. The role of instructional design in persuasion: A comics approach for improving cybersecurity. *International Journal of Human-Computer Interaction* 32, 3 (2016), 215–257.

A Codes

Based on the interview data, the research team developed this set of twenty low-level codes, grouped into higher-order themes:

Additional Information Requested

- How questions
  Example: *Just like how the barrier works, a little more detail.*
- What questions
  Example: *I would like to know exactly what information from medical records would be shown.*
- Who questions
  Example: *I don't know what the who the nonprofit is partnering with.*

Design Feedback

- Alternative presentations
  - Links to more detailed information
    Example: *I'd have a link or something to explain what general patterns means, what's the full detail, maybe as a side if they really were interested in knowing.*
  - Terms of use / consent documents
    Example: *I think I would definitely start with the thing that comes to mind first are informed consents that we sign as participants, and they're very clear about how will your data will be stored and who has access, how will it be de-identified.*
  - Video or animation
    Example: *What you could do is some sort of like animation type thing with a video-like format.*
- Icons
  - Color
    Example: *The green and red doesn't work for me.*
  - Locks
    Example: *I like the look of the lock.*
  - Privacy barrier
    Example: *A label of some sort beyond privacy barrier might be helpful.*
- Things people liked
  Example: *I like things that make it faster to read.*

Participant Understanding

- Did not understand
  Example: *So I'm not clear as to what the protection actually does.*
- Misconception
  Example: *It allows the people who utilize the information, say the law enforcement and medical professionals, it would allow them to share that information amongst themselves in a secure network without allowing the people who want to get that information to abuse that information.*
- DP as anonymization
  Example: *The only way I could explain it would be that an individual's personally identifying details would not be included with their medical records.*
- DP as fake data
  Example: *It's basically saying that we might put fake data in some parts of it.*
- Other PETs
  Example: *So basically there's like some kind of firewall that keeps my privacy safe.*
- User-generated metaphors
  Example: *It's kind of like an egg. You know, you crack it open and you don't know if it's going to be rotten inside or not. But I don't know what chicken it came from, so I can't blame the chicken.*

Reasoning About Data Sharing

- Benefits
  Example: *I actually think that people like data analysts or employee university employees probably want to see my information. Like in that case, that's when it's okay for privacy to be breached. Because it's for the purpose of the study.*
- Concerns
  - Concerns or skepticism about adequacy of protection
    Example: *It sounds good, but I just read too many things about the Internet not being so secure as we would like.*
  - Data disclosure risks (or lack thereof)
    Example: *Especially like insurance companies, I would want to make sure that it's not being shared without my knowledge.*
  - Lack of concern about privacy in general
    Example: *I don't care about my personal information being released.*

Sets of codes were also developed for the open-text survey responses. The following codes are related to respondents' reasoning about data sharing.

- Relationship with doctor
  Example: *I believe that if the doctors office is working with the non profit, I believe I trust them, there would also be massive repercussions if they were to do anything wrong with the records.*
- Want more info
  Example: *before i say yes, i would need more info such as-will they see my name, do they want my entire medical history, what kind of boundaries in medicine are they pushing and do they align with my beliefs and morals*
- Too risky or too private
  Example: *I think with all that's been going with abortion in the USA I'd be extremely wary of sharing medical data with a third party. Even if they have an extra layer of privacy protection they could still get hacked or the government could decide it has a right to that data.*
- Nothing to hide
  Example: *I would share my medical records with anyone who wanted to see them. This would not be an issue for me. I have nothing to hide.*
- Benefits of data sharing
  Example: *Yes, yes and absolutely yes. If this will help just ONE person who needs it, I would gladly share what I can to help them as long as my privacy was protected. Heck, even if it wasn't protected if it could still help then yes. I'm seeing commercials talking about wanting cancer Institutions to start doing this. This could have helped my dad perhaps. And if anyone would need to see his records to help others, I'd say yes.*
- Trust
  Example: *I want to help them with their research and I trust that they will be able to keep my information private.*
- Distrust
  Example: *i dont trust them*
- Money
  Example: *In todays society where information is money, I have a hard time trusting organizations or institutions with very private information such as medical records.*
- Frequency of data breaches
  Example: *Reassurances about security technology are hollow. Everything is breached eventually. It's just an arms race with the hackers.*
- Laws and Regulations
  Example: *Medical record information should be protected and private. That is what HIPPA is for.*
- Deletion
  Example: *Too loose in management, no note of when data will be deleted*

(which is the basic requirement for data collection in modern times), no mentions of security measures, no compensation for doing so nor any statement on how reputable the nonprofit organization is.

- Data already out there
  Example: *Because most medical information is public*

This final set of codes is related to feedback obtained through the online survey.

- Simplify
  Example: *Make it more simplified and shorter*
- More info about how protection works
  Example: *There needs to be more explanation about how the privacy protection works.*
- Other info requests
  Example: *need to be informed on where my data is going.*
- Positives
  Example: *I found the explanations of the privacy protection to be clear, concise, and easy to understand.*
- Confusion
  Example: *The picture is confusing to me. I don't understand why it needs two different sections*
- Nothing is foolproof
  Example: *Anything can be hacked. No one can be trusted*
- Distrust
  Example: *For me it's more of a feeling that I don't trust what is being presented as far as the safety of my information.*

B Survey Instrument

B.1 Instructions

In this survey we are going to ask you a series of questions about a hypothetical scenario. Please do your best to imagine yourself in this scenario and answer the questions as if you were actually making the decisions about which you will be asked.

B.2 Scenario Description

Imagine that during your next doctor's visit, your primary care doctor informs you that they are part of a non-profit organization trying to push the boundaries of medical research. The non-profit is asking patients around the country to share their medical records, which will be used to help medical research on improving treatment options and patient care. Your doctor, with your permission, can facilitate the non-profit getting the information they need.

B.3 Privacy Description

The non-profit organization will use an extra layer of privacy technology to protect your information. [Explanation inserted here.]

B.4 Comprehension Check

What kind of information does the non-profit want to collect? [Choice order randomized.]

- Medical records
- Music videos
- Book titles
- Location histories

B.5 Trust

Please indicate your agreement with the following statement: I trust the non-profit organization to protect my personal information privacy.

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree
- Prefer not to answer

B.6 Self-Efficacy

How confident are you that you have enough information to decide whether to share your medical record with the non-profit?

- Very confident
- Confident
- Moderately confident
- Slightly confident
- Not at all confident
- Prefer not to answer

How confident are you about deciding whether to share your medical record with the non-profit?

- Very confident
- Confident
- Moderately confident
- Slightly confident
- Not at all confident
- Prefer not to answer

B.7 Share

Would you be willing to share your medical record with the non-profit?

- Yes
- No
- Prefer not to answer

Please explain your decision. [Text entry.]

B.8 Objective Comprehension

For each of the following statements, please indicate if you expect the following to be true or false if you share your medical record with the non-profit.
An employee working for the non-profit, such as a data analyst, could be able to see my exact medical history.

- True
- False
- I don't know
- Prefer not to answer

A criminal or foreign government that hacks the non-profit could learn my medical history.

- True
- False
- I don't know
- Prefer not to answer

A law enforcement organization could access my medical history with a court order requesting this data from the non-profit.

- True
- False
- I don't know
- Prefer not to answer

Graphs or informational charts created using information given to the non-profit could reveal my medical history.

- True
- False
- I don't know
- Prefer not to answer

Data that the non-profit shares with other organizations doing medical research could reveal my medical history.

- True
- False
- I don't know
- Prefer not to answer

B.9 Thoroughness

Please indicate your agreement with the following statement: I feel that it was explained thoroughly to me how the non-profit protects patient privacy.

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree
- Prefer not to answer

B.10 Subjective Understanding

How confident are you in your understanding of the privacy protection?

- Very confident
- Confident
- Moderately confident
- Slightly confident
- Not at all confident
- Prefer not to answer

B.11 Feedback

What feedback (if any) would you like to share about the explanations of privacy protection? [Text entry.]

B.12 PETs

Have you ever heard of the following technologies? (select all that apply) [Choice order randomized.]

- Differential privacy
- End-to-end encryption
- Secure multi-party computation
- Deliquescent security
- None of the above
- Prefer not to answer

Which of these technologies do you think was described in the survey? [Choice order randomized.]

- Differential privacy
- End-to-end encryption
- Secure multi-party computation
- Deliquescent security
- None of the above
- Prefer not to answer

Please explain your reasoning. [Text entry.]

B.13 Background

How familiar are you with the following computer and Internet-related items? Please choose a number between 1 and 5, where 1 represents no understanding and 5 represents full understanding of the item. (Each item also offered 'prefer not to answer' option.)

- Advanced Search
- PDF
- Spyware
- Wiki
- Cache
- Phishing

In what year were you born? (four digits please) [Text entry.]
What is your gender? [Multiselect.]

- Man
- Woman
- Non-binary
- Prefer to self describe: [Text entry.]
- Prefer not to answer

Please specify your race/ethnicity (select all that apply).

- Hispanic, Latino, or Spanish
- Black or African American
- White
- American Indian or Alaska Native
- Asian, Native Hawaiian, or Pacific Islander
- Prefer to self describe: [Text entry.]
- Prefer not to answer

What is the highest level of school you have completed or the highest degree you have received?

- Less than high school degree
- High school graduate (high school diploma or equivalent including GED)
- Some college but no degree
- Associate's degree
- Bachelor's degree
- Advanced degree (e.g., Master's, doctorate)
- Prefer not to answer

Which of the following best describes your educational background or job field?

- I have an education in, or work in, the field of computer science, computer engineering or IT.
- I DO NOT have an education in, nor do I work in, the field of computer science, computer engineering or IT.
- Prefer not to answer

Which one of the following includes your total HOUSEHOLD income for last year, before taxes?

- Less than $10,000
- $10,000 to under $20,000
- $20,000 to under $30,000
- $30,000 to under $40,000
- $40,000 to under $50,000
- $50,000 to under $65,000
- $65,000 to under $80,000
- $80,000 to under $100,000
- $100,000 to under $125,000
- $125,000 to under $150,000
- $150,000 to under $200,000
- $200,000 or more
- Prefer not to answer

C Demographics

Table 4 describes the demographics of the 24 participants in the main interview study. Table 5 describes the demographics of the 10 participants who participated in the follow-up interviews. Table 6 summarizes the demographics of the survey respondents. Note that respondents could select multiple values for race/ethnicity and gender and that many respondents selected multiple options for race/ethnicity but did not explicitly describe themselves as multiracial. Table 7 displays the approximate percentage of respondents who expressed familiarity with various PETs out of all respondents who answered this question (n=684).

Table 4: Participant Demographics: Initial Interviews

Demographic Attribute		Count
Gender	Female	10
	Male	14
Age	< 20	2
	20-29	9
	30-39	6
	40-49	4
	50+	3
Race	Asian	1
	Black or African American	4
	Mixed, Multiracial, or Biracial	3
	White or Caucasian	16
Education	Secondary education (e.g. GED / GCSE)	1
	High school diploma / A-levels	11
	Technical / community college	4
	Undergraduate degree (BA / BSc / other)	5
	Graduate degree	2
	Doctorate degree (PhD / other)	1

Table 5: Participant Demographics: Follow-up Interviews

Demographic Attribute		Count
Gender	Female	5
	Male	5
Age	< 20	1
	20-29	3
	30-39	1
	40-49	1
	50+	4
Race	Asian	3
	Black or African American	1
	Mixed, Multiracial, or Biracial	2
	White or Caucasian	3
	Native American	1
Education	High school diploma / A-levels	2
	Technical / community college	2
	Undergraduate degree (BA / BSc / other)	5
	Doctorate degree (PhD / other)	1

Table 6: Respondent Demographics

Demographic Attribute		Count
Gender	Woman	343
	Man	335
	Non-binary	15
	Agender / Gender-fluid afab / genderqueer / they	5
Age	< 20	12
	20-29	249
	30-39	219
	40-49	98
	50+	119
Race/Ethnicity	Hispanic, Latino, or Spanish	83
	Black or African American	68
	White	478
	American Indian or Alaska Native	12
	Asian, Native Hawaiian, or Pacific Islander	110
	Multiracial or Mixed race	4
Education	High school or less	124
	Some college	233
	Bachelor's or above	337
Income	Less than $10,000	41
	$10,000 to under $20,000	53
	$20,000 to under $30,000	79
	$30,000 to under $40,000	68
	$40,000 to under $50,000	65
	$50,000 to under $65,000	85
	$65,000 to under $80,000	88
	$80,000 to under $100,000	51
	$100,000 to under $125,000	57
	$125,000 to under $150,000	30
	$150,000 to under $200,000	25
	$200,000 or more	32
Tech	Education or work in CSE/IT	148
	No education nor work in CSE/IT	527

Table 7: Familiarity with PETs

PET	#	%
End-to-end encryption	439	64%
Differential privacy	32	5%
Secure multi-party computation	26	4%
Deliquescent security (distractor)	3	<1%
None of the above	237	35%

D Designs

Table 8 lists all of the original metaphor texts. Figure 4 shows representative examples of our diagrams. Figure 5 shows our original privacy labels, and figure 6 shows the final versions. Figure 7 shows an example of the kind of Miro board that a participant in one of our follow-up interviews would have interacted with. Figure 8 shows how our designs evolved over time.

Table 8: Original Metaphor Descriptions

Local	Central
Sharing data with the protection of this technology is like donating a penny to a crowdfunding campaign. No one will know with certainty that you donated. The sum of the donations from a large group of people will be valuable to our data analysts.	*Publishing statistics, graphs, or tables using this technology is like publishing a blurry photo of the database that allows the viewer to see general patterns while hiding individual details. However, someone who obtained access to the database would be able to see all of the collected information in full detail.*
The technology works something like this: Imagine that we are collecting photographs, but instead of collecting the raw images, we blur the images, and only collect the blurry images, so that little is revealed about you as an individual. Anyone who accesses our collected data will only see the blurry images, rather than the originals.	*Publishing statistics, graphs, or tables using this technology is like publishing a photo of a mosaic, taken from a distance. People viewing this photo would not be able to see the individual tiles—in other words, individuals' data—yet they would still be able to see the overall picture. However, someone with direct access to the mosaic would be able to discern the individual tiles.*

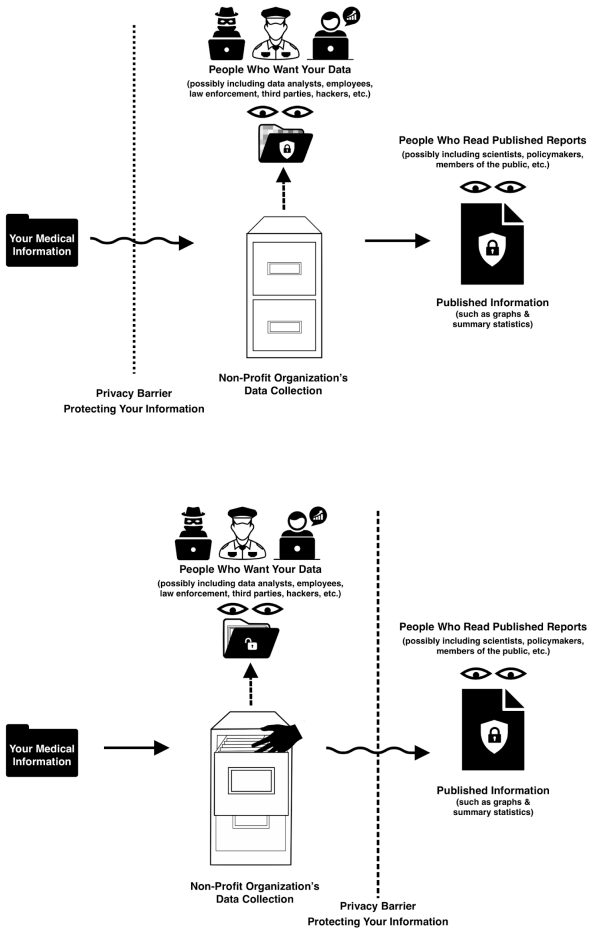


Figure 4: Top: Diagram for local model. Bottom: Diagram for central model.

Privacy protection
A person looking at graphs or informational charts created using information given to the non-profit will not be able to see your information.
A criminal or foreign government that hacks the non-profit will not be able to see your information.
A law enforcement organization with a court order requesting this data from the non-profit will not be able to see your information.
Employees, such as data analysts, working for the non-profit organization will not be able to see your information.
Other organizations doing medical research with whom the non-profit organization shares data will not be able to see your information.

Who Can See Your Data	Without Privacy Protection	With Privacy Protection
If people view graphs or informational charts created using information given to the non-profit...	...they might be able to see your information.	...they will not be able to see your information.
If hackers—like criminals or foreign governments— successfully attack the non-profit...	...they might be able to see your information.	...they will not be able to see your information.
If law enforcement with a court order requests your information from the non-profit...	...they might be able to see your information.	...they will not be able to see your information.
If employees of the non-profit organization, such as data analysts, access the organization's data...	...they might be able to see your information.	...they will not be able to see your information.
If organizations with whom the non-profit organization collaborates access shared data...	...they might be able to see your information.	...they will not be able to see your information.

(a) Local

Privacy protection
A person looking at graphs or informational charts created using information given to the non-profit will not be able to see your information.

Who Can See Your Data	Without Privacy Protection	With Privacy Protection
If people view graphs or informational charts created using information given to the non-profit...	...they might be able to see your information.	...they will not be able to see your information.
If hackers—like criminals or foreign governments— successfully attack the non-profit...	...they might be able to see your information.	...they might be able to see your information.
If law enforcement with a court order requests your information from the non-profit...	...they might be able to see your information.	...they might be able to see your information.
If employees of the non-profit organization, such as data analysts, access the organization's data...	...they might be able to see your information.	...they might be able to see your information.
If organizations with whom the non-profit organization collaborates access shared data...	...they might be able to see your information.	...they might be able to see your information.

(b) Central

Figure 5: Original Privacy Labels.

Who Can See Your Data	Without Privacy Protection	With Privacy Protection
Viewers of graphs or informational charts created using information given to the non-profit…	…might be able to see your information.	…will not be able to see your information.
Hackers—like criminals or foreign governments— who successfully attack the non-profit…	…might be able to see your information.	…will not be able to see your information.
Law enforcement with a court order requesting your information from the non-profit…	…might be able to see your information.	…will not be able to see your information.
Employees of the non-profit, such as data analysts, who work with the non-profit's data…	…might be able to see your information.	…will not be able to see your information.
Organizations collaborating with the non-profit that are given access to the non-profit's data…	…might be able to see your information.	…will not be able to see your information.

Who Can See Your Data	Without Privacy Protection	With Privacy Protection
Viewers of graphs or informational charts created using information given to the non-profit…	…might be able to see your information.	…will not be able to see your information.
Hackers—like criminals or foreign governments— who successfully attack the non-profit…	…might be able to see your information.	…might be able to see your information.
Law enforcement with a court order requesting your information from the non-profit…	…might be able to see your information.	…might be able to see your information.
Employees of the non-profit, such as data analysts, who work with the non-profit's data…	…might be able to see your information.	…might be able to see your information.
Organizations collaborating with the non-profit that are given access to the non-profit's data…	…might be able to see your information.	…might be able to see your information.

Figure 6: Final Privacy Labels.

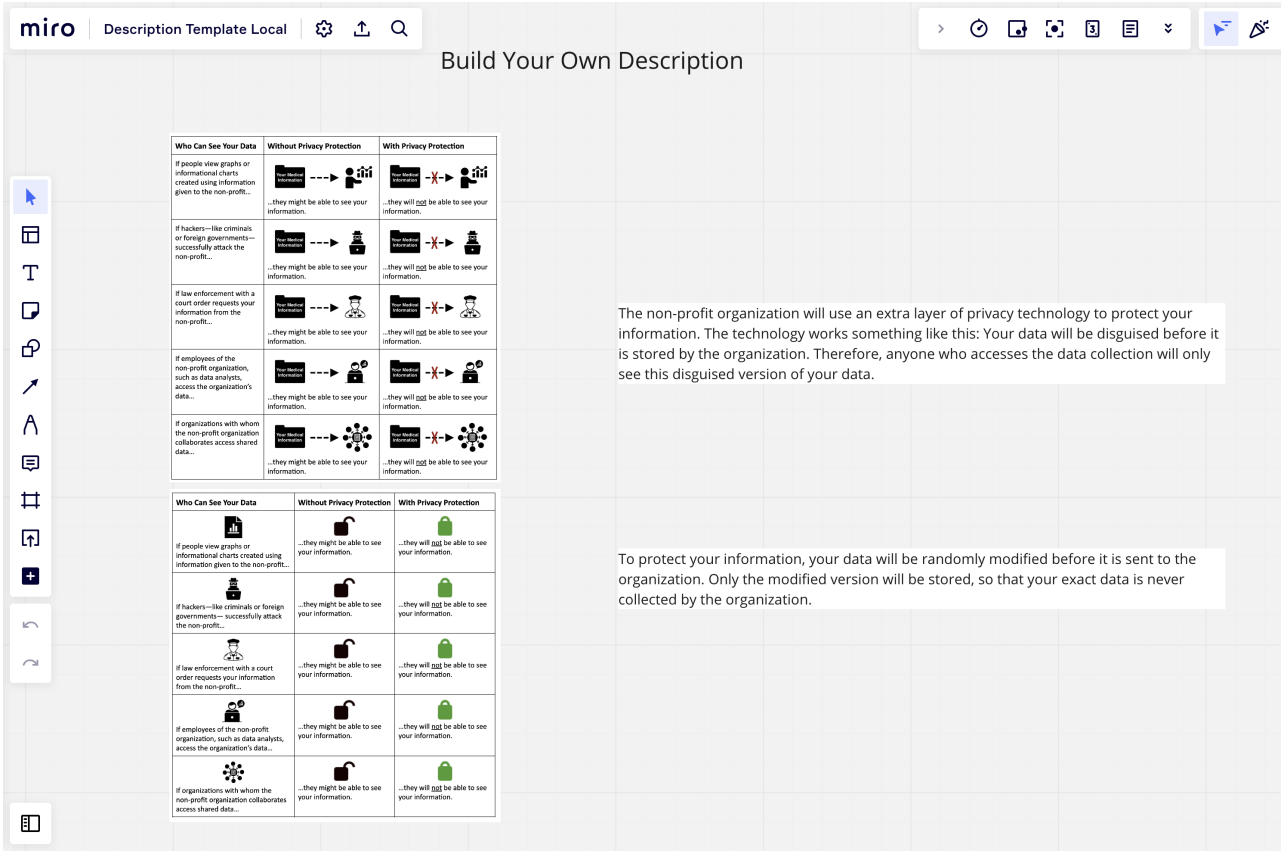


Figure 7: Example of the Miro board setup used for the follow-up interviews.

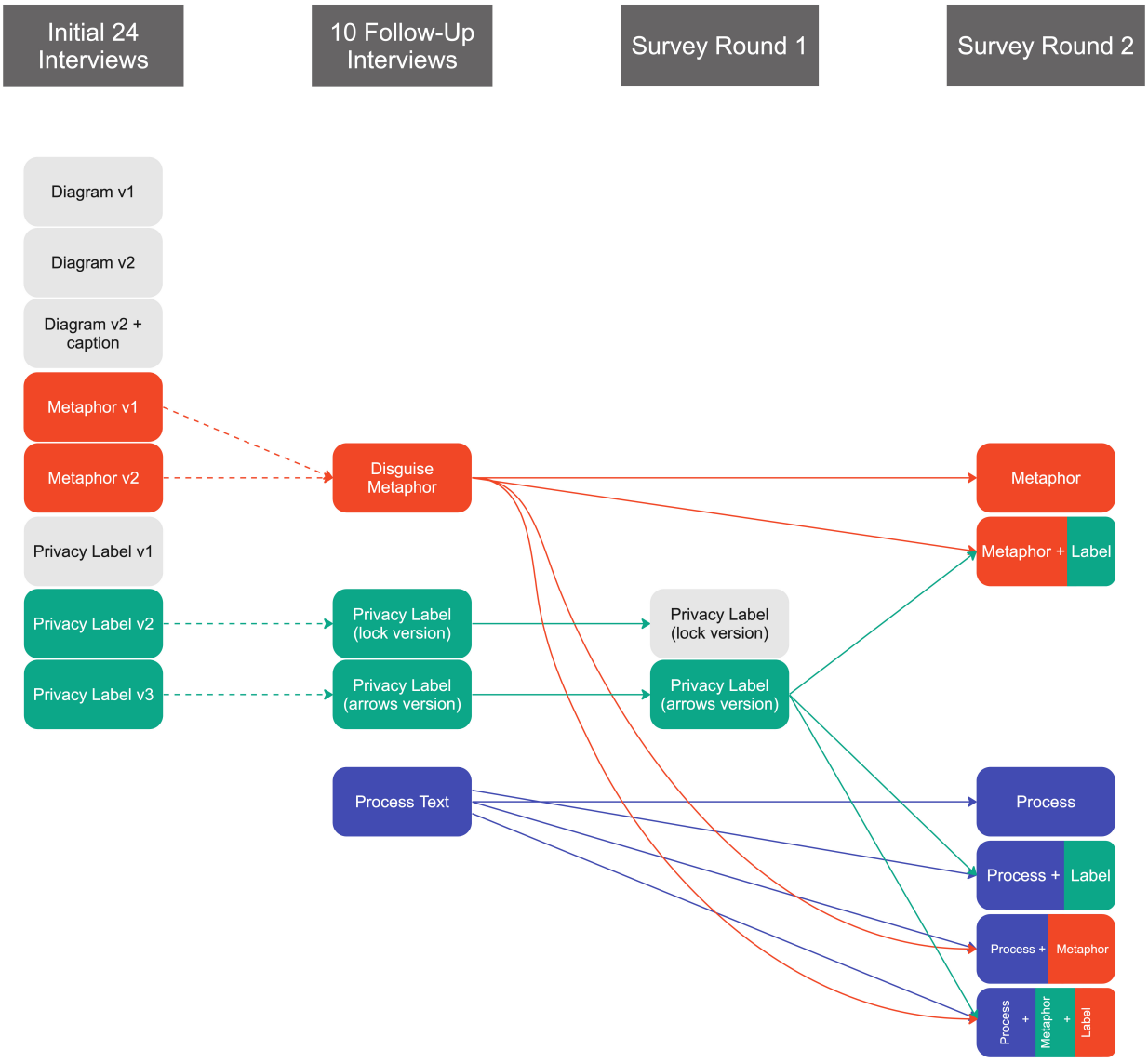


Figure 8: In the initial interviews, we evaluated multiple versions of each explanation type for both the local and central models. Based on participant feedback, we dropped the diagram explanations and modified the privacy labels. During the follow-up interviews, we developed a new metaphor and introduced a text with information about the data protection process. Next, we compared the two privacy labels through a survey, and dropped the version with locks. Finally, we evaluated the disguise metaphor, the process text, and combinations of these texts and the privacy label with arrows.

E Additional Tables

Table 12 reports results from the regression model for the data-sharing decision. Table 11 reports results from the regression models for objective comprehension and subjective understanding. Table 9 reports results from the regression models for trust, perceived thoroughness, and self-efficacy. Finally, Table 10 displays the proportion of respondents per condition who answered each comprehension correctly (+) and incorrectly (-). Since some respondents selected 'I don't know,' these percentages may not add to 1.

Variable	***Share***	
	OR	CI
Model: Local	**1.46***	[1.07, 2]
Expl: Metaphor	1.09	[0.57, 2.12]
Expl: Process	0.75	[0.38, 1.45]
Expl: Process+Metaphor	0.81	[0.41, 1.58]
Expl: Arrow Label	0.52	[0.26, 1.04]
Expl: Label+Metaphor	0.80	[0.4, 1.56]
Expl: Label+Process	1.33	[0.7, 2.56]
Expl: Label+Process+Metaphor	0.65	[0.33, 1.28]
Internet Skill	0.95	[0.79, 1.15]

Table 9: Results from regression model for data-sharing decision. We report odds ratios (OR) and corresponding 95% CIs. An OR > 1 indicates an increase in odds, while an OR < 1 indicates a decrease.

Model	Explanation	Hack		Law		Org		Graph		Share	
		+	-	+	-	+	-	+	-	+	-
Central	Metaphor	0.78	0.05	0.54	0.14	0.89	0.03	0.41	0.46	0.54	0.22
Local	Metaphor	0.26	0.47	0.24	0.45	0.37	0.37	0.39	0.42	0.26	0.47
Central	Process	0.50	0.28	0.50	0.17	0.52	0.17	0.42	0.32	0.45	0.32
Local	Process	0.28	0.48	0.3	0.50	0.38	0.42	0.32	0.28	0.28	0.52
Central	Process+Metaphor	0.79	0.11	0.76	0.05	0.92	0.08	0.61	0.34	0.50	0.37
Local	Process+Metaphor	0.36	0.38	0.33	0.33	0.44	0.41	0.49	0.31	0.38	0.41
Central	ArrowLabel	0.92	0.03	0.95	0.00	0.92	0.05	0.58	0.26	0.82	0.00
Local	ArrowLabel	0.51	0.32	0.44	0.37	0.46	0.32	0.63	0.24	0.63	0.27
Central	Label+Metaphor	0.85	0.13	0.82	0.05	0.82	0.08	0.64	0.26	0.77	0.15
Local	Label+Metaphor	0.72	0.22	0.58	0.33	0.58	0.33	0.69	0.19	0.61	0.31
Central	Label+Process	0.85	0.12	0.82	0.05	0.65	0.15	0.6	0.25	0.62	0.22
Local	Label+Process	0.59	0.28	0.44	0.31	0.67	0.21	0.62	0.18	0.67	0.28
Central	Label+Process+Metaphor	0.85	0.05	0.82	0.03	0.80	0.15	0.45	0.40	0.70	0.15
Local	Label+Process+Metaphor	0.56	0.26	0.59	0.18	0.56	0.26	0.69	0.15	0.49	0.26
Central	Xiong	0.91	0.03	0.44	0.12	0.62	0.15	0.35	0.44	0.44	0.32
Local	Xiong	0.33	0.31	0.23	0.44	0.33	0.41	0.56	0.23	0.41	0.46

Table 10: Accuracy of Privacy Expectations

Variable	***Objective Comprehension***		***Subjective Understanding***	
	β	CI	OR	CI
Model: Local	$\mathbf{-1.17}^{***}$	[−1.42, −0.92]	0.79	[0.61, 1.03]
Expl: Metaphor	0.09	[−0.45, 0.62]	1.63	[0.91, 2.90]
Expl: Process	−0.33	[−0.86, 0.19]	1.39	[0.79, 2.45]
Expl: Process+Metaphor	0.47	[−0.06, 1]	$\mathbf{1.79}^{*}$	[1.01, 3.18]
Expl: Arrow Label	$\mathbf{1.15}^{***}$	[0.62, 1.68]	1.35	[0.75, 2.42]
Expl: Label+Metaphor	$\mathbf{1.26}^{***}$	[0.73, 1.8]	1.51	[0.86, 2.67]
Expl: Label+Process	$\mathbf{0.97}^{***}$	[0.45, 1.5]	1.39	[0.79, 2.44]
Expl: Label+Process+Metaphor	$\mathbf{0.95}^{***}$	[0.43, 1.48]	1.48	[0.84, 2.6]
Internet Skill	$\mathbf{0.20}^{**}$	[0.05, 0.35]	$\mathbf{1.36}^{***}$	[1.16, 1.6]

Table 11: *Left:* results from linear regression models for objective comprehension. We report regression coefficients (β) and 95% CIs for these coefficients. $\beta > 0$ indicates an increase while $\beta < 0$ indicates a decrease. *Right:* results from ordinal regression models for subjective understanding. We report odds ratios (OR) and corresponding 95% CIs. An OR > 1 indicates an increase in odds, while an OR < 1 indicates a decrease. For both columns, we use the Xiong et al. explanation as the reference level explanation.* * $\mathbf{p}< 0.05$; ** $\mathbf{p}< 0.01$; *** $\mathbf{p}< 0.001$.

Variable	***Trust***		***Thoroughness***		***SE (Info)***		***SE (Confidence)***	
	OR	CI	OR	CI	OR	CI	OR	CI
Model: Local	$\mathbf{1.70}^{***}$	[1.3, 2.23]	1.08	[0.83, 1.41]	0.87	[0.67, 1.13]	0.90	[0.69, 1.17]
Expl: Metaphor	1.46	[0.82, 2.6]	1.28	[0.72, 2.27]	1.21	[0.67, 2.19]	1.65	[0.91, 2.99]
Expl: Process	0.99	[0.56, 1.73]	0.81	[0.46, 1.43]	0.70	[0.39, 1.23]	0.95	[0.53, 1.68]
Expl: Process+Metaphor	1.45	[0.81, 2.58]	1.55	[0.88, 2.74]	0.93	[0.52, 1.64]	1.19	[0.67, 2.11]
Expl: Arrow Label	0.85	[0.48, 1.52]	0.68	[0.38, 1.21]	1.15	[0.64, 2.05]	1.44	[0.8, 2.61]
Expl: Label+Metaphor	1.18	[0.66, 2.12]	1.73	[0.97, 3.09]	0.96	[0.53, 1.72]	1.16	[0.65, 2.06]
Expl: Label+Process	$\mathbf{1.97}^{*}$	[1.12, 3.47]	1.22	[0.7, 2.15]	1.08	[0.62, 1.87]	1.06	[0.61, 1.87]
Expl: Label+Process+Metaphor	0.94	[0.54, 1.64]	1.38	[0.79, 2.41]	0.98	[0.55, 1.73]	1.07	[0.6, 1.9]
Internet Skill	0.96	[0.82, 1.13]	1.06	[0.9, 1.25]	$\mathbf{1.21}^{*}$	[1.03, 1.41]	$\mathbf{1.28}^{**}$	[1.09, 1.5]

Table 12: Results from regression models for trust, perceived thoroughness, and self-efficacy, with the Xiong et al. explanation as the reference level explanation. Again we report odds ratios with 95% CIs. An OR > 1 indicates an increase in odds, while an OR < 1 indicates a decrease.