Lean Privacy Review: Collecting Users' Privacy Concerns of Data Practices at a Low Cost

HAOJIAN JIN, HONG SHEN, MAYANK JAIN, SWARUN KUMAR, and
JASON I. HONG, Carnegie Mellon University

Today, industry practitioners (e.g., data scientists, developers, product managers) rely on formal privacy reviews (a combination of user interviews, privacy risk assessments, etc.) in identifying potential customer acceptance issues with their organization's data practices. However, this process is slow and expensive, and practitioners often have to make ad-hoc privacy-related decisions with little actual feedback from users. We introduce Lean Privacy Review (LPR), a fast, cheap, and easy-to-access method to help practitioners collect direct feedback from users through the proxy of crowd workers in the early stages of design. LPR takes a proposed data practice, quickly breaks it down into smaller parts, generates a set of questionnaire surveys, solicits users' opinions, and summarizes those opinions in a compact form for practitioners to use. By doing so, LPR can help uncover the range and magnitude of different privacy concerns actual people have at a small fraction of the cost and wait-time for a formal review. We evaluated LPR using 12 real-world data practices with 240 crowd users and 24 data practitioners. Our results show that (1) the discovery of privacy concerns saturates as the number of evaluators exceeds 14 participants, which takes around 5.5 hours to complete (i.e., latency) and costs 3.7 hours of total crowd work (≈ $80 in our experiments); and (2) LPR finds 89% of privacy concerns identified by data practitioners as well as 139% additional privacy concerns that practitioners are not aware of, at a 6% estimated false alarm rate.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; • **Security and privacy → Human and societal aspects of security and privacy;**

Additional Key Words and Phrases: Privacy concern, data ethics, heuristic evaluation, privacy engineering

ACM Reference format:
Haojian Jin, Hong Shen, Mayank Jain, Swarun Kumar, and Jason I. Hong. 2021. Lean Privacy Review: Collecting Users' Privacy Concerns of Data Practices at a Low Cost. *ACM Trans. Comput.-Hum. Interact.* 28, 5, Article 34 (August 2021), 55 pages.
https://doi.org/10.1145/3463910

1 INTRODUCTION

Imagine a data scientist at Uber finds that users are more likely to accept surge pricing if their phone battery is low [22]. Incorporating this insight into the system may improve overall profits; however, it may also lead to negative headlines in the news media. Another example is the flawed launch of Google Buzz [18] in 2010, in which Google used users' contact information gathered from Gmail to generate their connections in a social network service. Today, industry practitioners face

Authors' address: H. Jin, H. Shen, M. Jain, S. Kumar, and J. I. Hong, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213.

1073-0516/2021/08-ART34 $15.00
https://doi.org/10.1145/3463910

a variety of similar dilemmas [2, 87, 111, 117, 134] in understanding how to best use potentially sensitive data in a manner that users will view as appropriate.

Navigating through these dilemmas requires significant expertise and effort [116]. Companies need to enumerate the range of problems pertaining to privacy, estimate the severity of each problem, and balance the potential fixes with competing interests (e.g., corporate revenue). Today, large companies like Google and Facebook have invested a great deal of resources in privacy. Privacy-related staff reside in both dedicated stand-alone teams and groups within other teams [13, 45, 88], providing decision-makers with tailored feedback through privacy reviews. A formal privacy review often involves multiple departments (e.g., legal, UX, business, public relations, and domain experts) and multiple rounds of discussion [4, 13, 28] (e.g., **privacy impact assessment** (**PIA**), user interviews, and **factorial vignette surveys (FVS)**). Such a review often requires a few weeks' turnaround time [13] and costs $10,000–$60,000 in human labor for companies [82].

This current approach to privacy reviews suffers from two major issues. First, since the privacy reviews are slow and costly, most small teams do not even have the resources to conduct such reviews. Even large organizations often only conduct comprehensive privacy reviews in a "sandwich" approach [13, 98, 119] (i.e., up-front specification followed by validation at the end). For many minor design decisions, practitioners can only offer their best guess (e.g., the "front page" test [99, 130]) as to users' expectations with little actual feedback from users [93, 136]. Second, experts conducting full privacy reviews often do not have the low-level technical knowledge of each data practice. The process requires practitioners to brief the data practice to experts and wait for experts' feedback to validate the final implementation, resulting in a considerable overhead to the whole organization [13].

This article presents **Lean Privacy Review** (**LPR**), a new discount privacy assessment method that can collect direct feedback from users at a low cost and provide tailored feedback for a data practice design, without requiring an actual technology implementation. At the heart of LPR is a novel approach that guides crowd workers in **actively** examining a specific data practice for privacy concerns, which reduces the cost and required resources for privacy concern inspection. Figure 1 presents an overview of the workflow of LPR. LPR asks each crowd worker to examine a description of data practice and express their concerns in **free text**, rather than only through passive responses to Likert questions (i.e., FVS) [12, 112]. Although individual crowd workers may have limited privacy expertise and only find few privacy problems, our experiments (Section 9.3.1) show that collectively aggregating their privacy concerns can achieve good and consistent coverage.

The design and development of LPR were deeply influenced by lean startup[1] [105], agile software development [23], discount usability testing [89, 91], and speed dating [32]: small teams do short development cycles with progressively refined prototypes, getting customer feedback at each step. We envision that LPR would enable two types of usages. First, practitioners can use LPR to conduct discount privacy reviews when a formal privacy review is not available. Second, the low-cost nature of LPR makes it possible for practitioners to iterate their data practices based on fast privacy feedback, as well as explore a broader set of alternative data practices in the early stages of design. Here, LPR is not designed to replace full privacy reviews, but rather serve as a

[1]We borrow the term "lean" from the concept of "lean startup" since LPR resembles two important properties of a lean startup. First, LPR collects direct feedback from users to help uncover their actual privacy concerns. Second, the privacy concern collection process only involves crowd workers and practitioners, making it more agile than a professional privacy review.

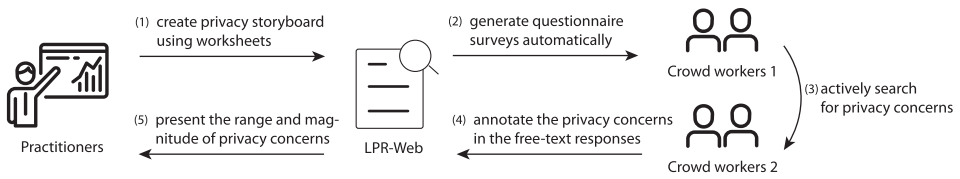


Fig. 1. The workflow of a discounted privacy review. (1) Practitioners first transform an initial data practice design into a privacy storyboard (Figure 2, left) using LPR's data action worksheet (Appendix B), and (2) LPR web system (LPR-Web) would generate a set of questionnaire surveys (Figure 2, middle) using a survey template. Practitioners then (3) send surveys to crowd workers and ask them to actively search for privacy concerns. (4) LPR then forwards the free-text responses to another group of crowd workers and asks them to label the privacy concerns using a taxonomy of privacy concerns (Figure 2, right and Appendix A). (5) Finally, LPR aggregates users' opinions and summarized the range and magnitude of privacy concerns in a compact form. In our evaluations, we found that practitioners can create the storyboard in 20–30 minutes, and the crowd review takes around 5.5 hours to complete (i.e., latency) and costs 3.7 hours of total crowd work ($\approx$\$80).

low-cost and complementary method to full privacy reviews, similar to how heuristic evaluation is complementary to full user studies.

Yet, asking crowd workers to actively examine a data practice is challenging. First, we lack an effective way to communicate data practices to non-tech-savvy users. For example, past work [46] has found that most non-specialists do not understand the details of the Facebook Cambridge Analytica data scandal [49, 125] even after being exposed to massive press coverage. Second, we need an approach to scaffold a privacy design review for non-tech-savvy users. Conducting a formal privacy review often takes hours for an expert. However, most users may not have the expertise to articulate a privacy problem and may be unable to engage for a long time to do so. Finally, interpreting the free-text responses is not easy for practitioners, who may lack the time and the expertise to analyze the qualitative responses. The rest of this article describes our solutions (Sections 4–6) to these key challenges in making LPR practical and a web-based system (Section 7) to support LPR.

1.1 A Usage Scenario of LPR

A data scientist (i.e., the data practitioner) in a search company wants to run an A/B test to determine which color is the best for web hyperlinks [7, 55]. On one hand, the data scientist feels that this is a rather innocuous experiment, especially since users will not face any direct financial loss. On the other hand, the data scientist worries that this experiment does not explicitly obtain users' consent, and some users may feel manipulated. Figure 1 illustrates her interaction with LPR to evaluate the privacy concerns of her proposed data practice design.

Privacy storyboarding (Section 4). LPR provides a worksheet (Appendix B) and a web interface (Figure 8) to help data practitioners think through a planned data practice and generate a privacy storyboard. A privacy storyboard contains a set of short free-form textual descriptions (see examples in Figure 2) organized in multiple data flows. These descriptions should contain all the necessary privacy-salient information (e.g., how the data are being collected, who it is being shared with, what information is being derived, and how the data are being used) to evaluate the privacy problems while being understandable by a general audience after a quick read. The worksheet provides detailed steps to help a data practitioner break a data practice down into smaller parts and generate the free-form textual descriptions. The web interface supports the data practitioner to organize these descriptions in data flows.

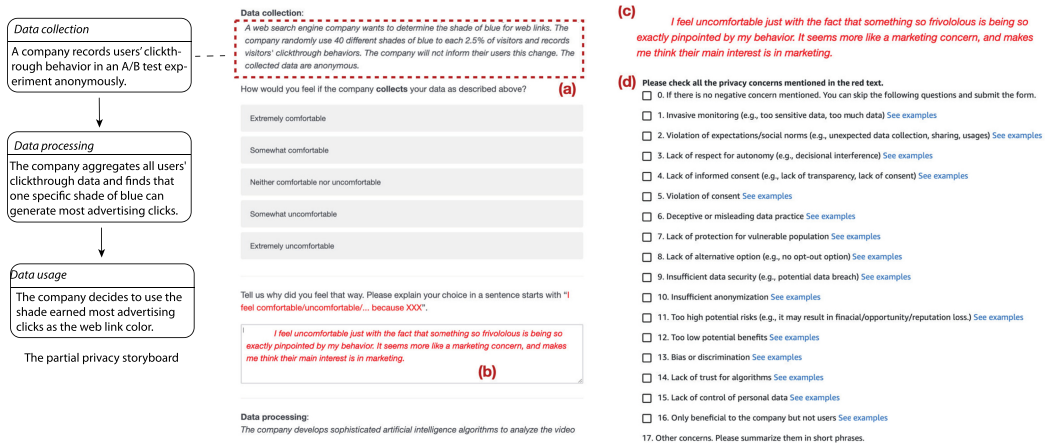


Fig. 2. An example use of LPR based on "Google's 41 Shades of Blue A/B test" [7, 55]. Suppose a practitioner wants to evaluate the privacy concerns of running an A/B test to determine which color is the best link color. She starts by creating a privacy storyboard (left). LPR then involves crowd workers in two different tasks: expressing privacy opinions (middle) and annotating free-text responses (right). LPR first asks crowd workers to examine a data action description (a) in free text (b), then forwards the collected free-text responses (c) to another group of workers for privacy concerns annotation (d).

Collecting direct feedback from users (Section 5). LPR-Web then generates a set of questionnaire surveys automatically based on the privacy storyboard, and distributes them to crowd workers. Each survey asks a crowd worker to express her concerns regarding each textual description and elaborate on the reasons in free text (Figure 2, middle). Each crowd task takes 5–15 minutes to complete. Similar to heuristic evaluation [91], a single individual might not find all the privacy concerns, but aggregating multiple crowd workers can improve the effectiveness since different people often discover different concerns.

Presenting feedback to practitioners (Section 6). Quantitatively interpreting users' free-text responses is challenging since users' concerns on privacy matters are often noisy and diverse [135]. LPR sends collected free-text responses (e.g., the red text in Figure 2(c)) to another group of crowd workers for annotation of privacy concerns (Figure 2, right). By doing so, LPR can aggregate free-text responses into various quantitative privacy spectra (see examples in Figures 9 and 10) to help practitioners understand the range and magnitude of users' privacy concerns.

1.2 Evaluation

We demonstrate through our evaluations that LPR is inexpensive, fast, consistent, and can help practitioners find potential concerns they are not aware of.

- **The applicability of privacy storyboarding** (Section 8). We selected 12 diverse real-world privacy stories and asked three participants to create privacy storyboards. Our results suggest that practitioners can generate such storyboards in a fast manner: each storyboard creation takes between 20 and 30 minutes.
- **The cost and latency of crowd inspection** (Section 9). LPR automatically generates surveys based on the user-generated storyboards and solicits crowd workers' opinions. We queried 20 crowd workers for each story. Our results suggest that the privacy concerns found by crowd workers converge. On average, aggregating the inspection results from any 14 crowd workers can uncover 97% of privacy concerns identified by all participants. The

process costs 3.7 hours of total crowd work (≈ $80 in our experiments[2]), and the requester needs to wait for 5.5 hours[3] to get the results.

— **The quality of inspection results** (Section 9). Three trained privacy researchers compared the crowd worker results with the results from 24 software and data practitioners (a mix of trained UX researchers, privacy engineers and software engineers). They found that LPR discovered 89% of privacy concerns identified by data practitioners as well as 139% additional privacy concerns that practitioners did not identify, at a 6% **false alarm** (**FA**) rate.

1.3 Contributions

Our specific research contributions are as follows:

— LPR is the first method that inspects privacy concerns of data practices using the crowd. LPR takes a proposed data practice, breaks it down into smaller parts, generates a set of questionnaire surveys, solicits users' opinions, and summarizes the opinions in a compact form for practitioners to use.

— We propose a new storytelling technique, privacy storyboarding, to help practitioners communicate a data practice and relevant privacy considerations without having to implement a system. We identify a set of simple and expressive vocabulary for data actions (i.e., data collection, sharing, processing, and usage) and organize these actions in a tree-like topology (see Section 4.3 and examples in Figures 4, 8, and 23–34). The tree topology makes the storyboard modular, so practitioners can incrementally modify and append more data applications/actions.

— We introduced 12 real-world data uses cases into data ethics discussion and evaluated LPR using these practices with 240 crowd users and 24 data practitioners. Our results show that LPR is inexpensive, fast, consistent, and can provide high-quality privacy review results.

2 RELATED WORK

LPR builds on ideas from four different areas: (1) privacy review by experts, (2) user research methods in understanding users' privacy concerns, (3) privacy surveys, and (4) crowdsourcing.

2.1 Privacy Review by Experts

Modern privacy regulations (e.g., [19]) argue that organizations should think about privacy in a proactive manner (i.e., Privacy by Design) rather than a reactive one (traditional privacy regulations). As such, one best practice before deploying a data practice is to have product teams consult internal privacy experts to vet problems beforehand [13], to make sure that the data practice is legal and poses an acceptable risk for the company's business [94]. These privacy experts are often a mix of experts from different relevant areas with specialized training in privacy, such as lawyers, UX researchers, privacy engineers, software engineers, and product designers [28], and might reside in dedicated stand-alone teams (e.g., a privacy council) or groups within the product team [88].

There exist a few frameworks for reviewing the privacy of a data practice, such as NIST Privacy Risk Assessment Methodology [93], PIA [136], and **Data Protection Impact Assessments** (**DPIA**) [95]. These frameworks are derived from information security research and treat privacy problems in a manner similar to security risks. For example, PRAM first asks experts to map out

[2]The cost includes the payment to both the crowdsourcing platforms and crowd workers.

[3]The latency is because crowd workers do not start the task immediately after tasks are published, and some workers may abandon the tasks.

the data processing pipeline within the target system, and then catalog contextual factors and data actions.[4] Experts then enumerate all potential problems associated with each data action and then assign scores to the likelihood and severity of each problem. In this way, builders can quantify and prioritize privacy risks for the organization (e.g., revenue loss from customer abandonment), and determine appropriate resource allocations to address the risks. The quality of such a privacy review is heavily dependent on the experts' expertise level.

Another major approach is guidelines-based, where privacy experts use a set of guidelines to design and analyze a specific data practice [57]. For example, the **Fair Information Practice Principles** (**FIPPS**) enumerates a set of guidelines—such as Transparency, Purpose Specification, and Data Minimization—that an ideal privacy-sensitive system should satisfy. Data protection authorities can use the FIPPS to regulate practitioners regarding how, when, and for what purpose data can be collected, used, and disclosed [57]. Nissenbaum's "decision heuristic" [47] includes a series of guidelines articulated in nine steps, which can be used to evaluate the "system or practice in question."

However, both the privacy engineering and guidelines-based approaches look at privacy primarily from the organization's perspective (i.e., the risks for their business) rather than the end-users' view (i.e., end-users' perceptions and concerns). The likelihood and severity scores are also best guesses from these experts, which might be different from actual users' perceptions of the privacy problems. Through LPR, users' expectations of the transmission of their personal information—once carefully structured and aggregated—have the potential to serve as an important resource for *bottom-up* approaches to privacy decision-making.

Furthermore, these privacy frameworks require expertise in privacy, and privacy experts are scarce and expensive [28]. This constraint has a two-fold effect in practice. First, smaller teams often do not have the resources to hire such specialists. Second, even for companies with dedicated specialists, having a dedicated team reflect on each data practice's potential privacy concerns is infeasible. For example, companies like Facebook run over a thousand data science experiments each day [41, 64]; however, even a lightweight privacy discussion involving only a few data scientists, developers, and product managers can cost the company several thousand US dollars [83, 124]. LPR is designed to be complementary to these methods, when privacy experts are not available. Further, the fast and low-cost nature allows practitioners to explore more alternative data practice designs, and collect more direct feedback from users throughout the development life cycle.

2.2 User Research Methods in Understanding Users' Privacy Concerns

Beyond the privacy review from a company's perspective, **Human computer interaction (HCI)** and privacy researchers have developed various methods to understand users' privacy concerns, such as surveys [6, 40, 127, 135], focus groups [33, 68, 70], interviews [66, 85], experiments [39], contextual inquiry [9], and analyzing online corpus [85]. However, existing methods suffer from two constraints.

First, studying users' privacy concerns through these methods are often costly in terms of both time and money. For example, to investigate regrets associated with users' posts on Facebook, Wang et al. [127] had 569 users participate in interviews, user diaries, and online surveys. The entire process cost over $1,000 in recruiting participants, and took the authors more than seven months to finish the studies. Naturally, most companies do not have the resources to justify such an investment and cannot wait for that long period. Instead, LPR aims to offer a low-cost and fast method that can help data practitioners collect users' direct feedback.

[4]Data actions are information system operations that process personal information [93].

Second, most UX studies (e.g., [127]) for data practices are conducted in a retrospective manner (i.e., when the target system is already deployed), since users often need to try the deployed system to understand their experiences. While techniques like Wizard of Oz [72] and similar alternative technologies [66] can help elicit users' generic privacy concerns, these privacy concerns often can only provide little direct feedback to tailored data practices. For example, Krombholz et al. [66] used Google Glass as an example of wearable technology to provoke participants' generic privacy concerns about wearables. However, users' privacy concerns regarding a data practice vary as the nuances of the data practice change [110, 135], and it is hard to emulate these nuances through an approximate set-up. In contrast, LPR develops a storytelling framework to help practitioners communicate their data practice without implementation. Further, LPR takes a procedural perspective, allowing users to dive into the procedure of the data practice to understand the nuances and inform practitioners which part of the process might be problematic.

2.3 Privacy Surveys

Privacy surveys are a common approach to measure users' privacy attitudes and concerns [12, 48, 48, 80, 123] across times [5, 67] and demographics [25, 30, 128]. One notable example is the over 30 privacy indexes created by Dr. Alan Westin [67]. These privacy indexes cover both the general level of privacy concerns of the public as well as the attitudes about specific privacy-related topics, for example, confidence in organizations that handle personal information, acceptance of a national identification system, and use of medical records for research.

A major limitation of these surveys is that participants' broad, generic privacy attitudes do not match well with context-specific, privacy-related behaviors, either actual or intended [26, 78]. Barkhuus [9] questioned the viability of obtaining universal answers in terms of people's "general" privacy practices, and argued for the use of more specific vocabularies to analyze contextually grounded privacy issues. A number of studies [6, 60, 133] have used the Contextual Integrity [92] framework to construct context-related questions that can help identify established privacy norms. For example, Apthorpe et al. [6] examined a range of settings, devices, and information types in specific contexts through questions like: "A sleep monitor records audio of its owner. How acceptable is it for the monitor to send this information to [different recipients]?" Contextual privacy surveys can provide valuable insights to establish privacy norms/guidelines. Still, they are often insufficient to provide direct feedback to a specific data practice, since these parameterized descriptions cannot capture the richness of complex data actions across multiple stakeholders.

The surveys generated by LPR fall into another group of privacy surveys, namely privacy incident surveys [29, 65, 135], which collect users' responses toward a specific scenario, aiming to capture the details of privacy-sensitive contexts. The most relevant work is ethical-response surveys [110], in which Schechter et al. [110] ran 3,539 FVS to study five scenarios and found that a minor change to the Facebook emotional contagion experiment design can reduce users' disapproval and concern significantly.

However, exploring alternative benign variants is challenging for most practitioners. For example, to study users' privacy concerns through FVS, experimenters need to carefully design the controlled variables across variants, collect a large number of responses, and analyze the results to illustrate statistical significance. Therefore, FVS is too expensive, in terms of time, cost, and expertise requirement, when data practitioners need a tool to support their decision-making in the day-to-day work. Our key insight into the process is that non-tech-savvy participants can contribute more than just answering multiple-choice questions. Instead, LPR asks participants to **actively** examine a specific data practice for privacy concerns and express their concerns in free-text. This design reduces the cost of privacy surveys for two reasons. First, by asking

participants to actively search for privacy concerns, experimenters do not need to control variable across surveys. Second, the free-text responses are more effective in collecting participants' privacy-related opinions than the Likert scores. A few free-text responses can cover multiple privacy concerns, while it may take hundreds of numerical scores to illustrate the statistical significance.

2.4 Crowdsourcing

Crowdsourcing is a widely used method for many tasks, such as gathering data to train algorithms [34], running user studies [62], proofreading and text editing [11], real-time captioning [69], writing news articles [63], creating taxonomies [21], penetration tests [37], exploring security configurations [58], and usability testing [75]. To the best of our knowledge, LPR is the first work to measure privacy concerns through crowdsourcing.

Crowdsourcing brings several unique benefits for the privacy review tasks, such as easy-to-access participants pool, low cost, fast response, and programmable demographics [75]. However, crowd workers may lack the motivation and expertise to engage in complex tasks and provide high-quality feedback [62].

We applied the insights from the prior crowdsourcing research to address these challenges. LPR offers a specific workflow, breaking a single privacy review task into two small types of tasks: expressing free-text privacy opinions and annotating free-text responses using a taxonomy of privacy concerns. We were inspired in part by the Find-Fix-Verify pattern in Soylent [11], which helps control the quality of crowdsourcing results. Kittur et al. [62] recommend making crowdsourced tasks more laborious to make it hard to cheat, so LPR asks participants to respond with more than 100 characters. MicroTalk [38] finds that argumentation, which requires workers to justify their responses, can improve results accuracy by 20%. Similarly, in addition to rate their comfortable level, LPR asks participants to elaborate on the reasons behind their rating.

3 DESIGN GOALS, CHALLENGES, AND DEVELOPMENT METHOD

In this section, we discuss the design goal, challenges and brief the development process of LPR.

3.1 LPR Design Goals

To recap, existing solutions to review privacy design issues suffer from a number of limitations, such as the lack of feedback from potential users [93, 136], high cost in terms of both time and resources [4, 13, 93, 110, 136], retrospective nature that are only useful once a system is already deployed [136], a slow review process (i.e., feedback latency) [110], and the ability to only offer generic feedback [6, 93]. In response to these limitations, we define three properties an ideal LPR solution should embody.

- **Human-centered privacy measurement.** The technique should collect users' privacy opinions directly and use these opinions to help practitioners identify privacy design flaws.
- **Low cost, light-weight, and accessible.** Existing review techniques (e.g., PRAM and FRS) are expensive and time-consuming. McQuinn et al. [82] estimate that a privacy audit can cost $10,000–$60,000. The audit process may take multiple weeks since it often involves multiple stakeholders (domain experts, UX researchers, privacy engineers, and executives) [13]. An ideal method should be fast, accessible, and low cost, so practitioners can collect users' direct feedback quickly and iterate the data practice design accordingly.
- **Practical feedback for practitioners.** The technique should provide concrete and quantitative feedback to data practitioners, not only which part of the data practice might be wrong, but also how severe the problem is (i.e., the percentage of affected people and

the severity of the impact). In this way, practitioners can test different variations to balance privacy-relevant design choices with opposing interests.

3.2 LPR Challenges

At the heart of LPR is a novel approach that guides non-specialists in **actively** examining a specific data practice for privacy concerns, which reduces the cost and required resources for privacy concern inspection. We develop and test LPR with crowd workers since they are easy to access for most practitioners. Besides, most crowdsourcing platforms allow requesters to specify the eligibility criteria so that practitioners can approximate the demographics of participants to match the target users group.[5] Along with these exciting benefits also come three significant challenges.

- **Communicating the privacy design to crowd workers.** Understanding the details of a data practice requires significant expertise and technical background.
- **Scaffolding the privacy design review.** Asking crowd workers to review a privacy design is challenging for two reasons: task duration and expertise. When a privacy engineer reviews the data practice, she needs to enumerate the potential privacy problems and then to assess the impact and likelihood. This process can take hours to complete; however, most crowd workers can hardly engage for such a long time. Besides, crowd workers may only have a vague understanding of privacy and may be unable to articulate the problem. For example, feedback such as "I just do not like this data practice" offers rather limited insights for practitioners.
- **Presenting the output to practitioners.** Interpreting results from crowd workers is not easy. First, practitioners may not have the time to read through all the raw feedback from crowd workers. Second, users' responses to privacy designs can be diverse and conflict with one another. As such, quantitatively assessing the results can be challenging.

3.3 A Summary of the Development Process

Over a span of two years, we designed, implemented, and evaluated the solutions for the above challenges iteratively. We started with a formative study (Section 4.1) to explore the design space of privacy storytelling, and used the obtained insights (Section 4.2) to devise an initial privacy storyboarding technique. We then coded real-world data practices using the proposed storytelling technique and iterated on the storyboard representation as we expanded the supported use cases (100+). In the end, we identified a set of simple and expressive vocabulary for privacy storyboarding (Section 4.3) and developed a worksheet (Appendix B) to facilitate data action analysis (Section 4.4).

We then iterated the design of privacy inspection tasks for non-specialist crowd workers in seven rounds of experimentation, aiming to communicate a privacy storyboard through a reusable template and collect participants' responses in an aggregatable manner. In each round, we experimented with different task designs and sent at least 50 surveys (5 storyboards x 10 participants) to crowd workers. The authors then manually went through the responses, analyzed the inspection quality, potential confusion, as well as the quantitative metrics (e.g., task completion rate, and duration), and used these feedback to inform the task design in the next round. To avoid the learning effect, all the crowd workers can only participate in the study once. The required completion times for different tasks vary; we paid participants an average hourly rate of $15, by offering an extra bonus after survey completion.

[5]LPR should perform the best when the inspection participants represent the target users group. However, crowd workers may not resemble the target user population. We discuss this limitation in Section 11.2.

Our initial intuition was to apply the design of usability heuristic evaluation to privacy problems (Section 5.1). We derived heuristics from past privacy research literature, and asked crowd workers to inspect the privacy problems of each data action using offered heuristics. However, the heuristic-based approach suffered from a few importation limitations. We focused on the iteration of heuristics in the first four rounds and then shifted to a two-stage design (collecting free-text responses (Section 5) first and then asking crowd workers to annotate (Section 6)). Through the process, we refined the storytelling techniques to communicate the details better, formulated the privacy problems as privacy concerns, articulated the design of the crowd privacy inspection tasks.

4 COMMUNICATING THE PRIVACY DESIGN: PRIVACY STORYBOARDING

Communicating a data practice between non-specialists and data practitioners is challenging. Existing solutions tend to choose between either being easy-to-understand for non-specialists (e.g., UX storyboards) or easy-to-generate for data practitioners (e.g., context data flow diagram [77, 114]). This section describes the design and development of the privacy storyboarding technique we created for LPR. Here, our goal is to provide a set of simple and expressive vocabularies to help practitioners effectively communicate their data practices to non-specialists without having to fully implement a working system.

4.1 Formative Study

To explore the design space of privacy storytelling, we first conducted semi-structured interviews with six participants (three male and three female, mean age = 25.3, max = 30, min = 21) to understand the communication challenges as well as informing the initial design of LPR. We recruited these participants through the flyers posted on a university campus. Two of them were professionals who live around the campus, and the rest were students. None of the participants had received specialized privacy training previously.

We chose six privacy incidents[6] (see Table 1) and organized the privacy scenarios similar to Woodruff et al. [135]. Namely, we identified the context, procedure, and the privacy-relevant outcomes for each privacy scenarios explicitly. We then tested each scenario with three storytelling techniques from the literature.

(1) **A long textual paragraph** [110, 113] (Figure 3, left). Scenario-based text descriptions/ fiction/ narratives have been commonly used in many privacy-related studies (e.g., [110, 113, 135]). We chose this approach as the baseline.
(2) **A hand-crafted privacy storyboard** (Figure 3, right). Using a graphical representation to describe a privacy practice is inspired by scenario-based storyboarding in UX design [107, 121]. Designers often create storyboards to help users perceive, interpret, and make sense of proposed functionalities. Graphical representations are more engaging and less biased while the use of words may bias a user's reaction to particular technologies [122].
(3) **Multiple short story snippets** [109]. The design of multiple short snippets is inspired by the survey on the Ethics of Scientific Experimentation [109], in which Schechter et al. listed the procedures of a data practice in bullet points. Such a design has two advantages. First, users' privacy opinions are often beyond the utilitarian perspective. Even if the outcome of the data practice is highly beneficial, other procedural factors, such as deception, even with goodwill, may irritate users. Second, we want to help practitioners understand which step in their data practice is problematic and why, i.e., locate the problem in a specific step of the data practice.

[6]We discuss the scenario selection criteria in Section 8.1.

Table 1. Privacy Scenarios Used in the Formative Study

#	Original source	Abstract
1	Target Pregnancy Prediction (Figure 3) [53]	A retail company develops a customer tracking technology that can predict if a female customer is likely pregnant.
2	Facebook Emotion Contagion study [65]	A social network company conducts a massive psychological experiment, manipulating their news feeds to assess the effects on their emotions.
3	Google 41 Shades of Blue [7, 55]	A search engine company runs a large scale A/B test experiment, randomly showing 41 different shades of blue to visitors to determine the best color of links.
4	Cambridge Analytica Scandal [49, 125]	A political consulting firm harvests the personal data of millions of people's social network profiles without their consent and used it for political advertising purpose.
5	OkCupid match score manipulation [54, 134]	A mobile dating app manipulates customer data to see how users of its dating service would react to one another.
6	Device-based price discrimination [2]	An online travel agency shows different prices to users with different devices (e.g., mobile vs. PC, Android vs. iOS).

For (1) and (2), we first presented the whole scenario, and then asked the participants: "If someone you cared about were a candidate participant for this data practice, would you want that person to be included as a participant?"[7] For (3), we presented each snippet one by one and asked the same question separately. Participants were also asked to explain their rationale for each question. To trigger participants' reflection and test their comprehension, we also presented a few variations for each scenario. For example, in the Target Pregnancy Prediction [53], we asked participants, "What if the retail company wants to predict if users are sports fans and send beer coupons to them during the World Cup?"

The experimental design is similar to Jin et al. [61]. Each participant went through all six privacy incidents in a randomized order, and we randomly assigned one of the three storytelling techniques to each incident. So we tested each storytelling method in 12 tasks (6 incidents x 2 participants). Each interview lasted around one hour. We compensated each participant $20 for their time. None of the participants were aware of these incidents before the study.

4.2 Findings and Design Choices

This pilot study resulted in the following findings.

Graphical vs. Textual representation. Graphical representations for privacy storytelling suffer from several limitations. First, generating a visual storyboard is challenging. The authors need to have a thorough understanding of the data lifecycle, decent visual design skills, and familiarity with UX Storyboarding Guidelines (e.g., number of frames, level of detail, and the inclusion of text [121]). Second, manipulating the variables in visual storyboards is not easy. Visual storyboards are often developed to test distinct features. In contrast, privacy practices are often subtle and need multiple minor variations to probe the salient issues. For example, Schechter et al. [110] created ten variants for the Facebook Emotion Contagion study [65], each describing a small experiment change.

[7]This question is adapted from Schechter and Bravo-Lillo [110].

A social media company collected information about its users, including the posts its users shared online.

Using automated algorithmic model, it can measure whether users' posts were of a positive or negative mood.

Participants will not be identified and will remain anonymous.

In order to test whether the mood of a user's friends' posts will influence this user's mood, the company either randomly excluded some fraction of friends' negative posts [or positive posts] each time the news feed was loaded.

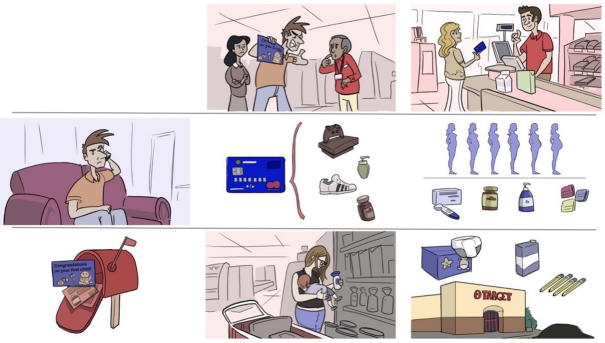


Fig. 3. Story telling techniques used in the formative study. Left: the plain text description of "#2 Facebook Emotion Contagion Study" [65]. Right: the storyboard representation of "#1 Target Pregnancy Prediction" [53].

Resulting design choice: We decided to use textual representation since it is easy to generate and iterate. We discuss the potential bias caused by the use of words in Section 12.3.

Multiple short snippets vs. a long paragraph. Short snippets can elicit more explicit privacy opinions from participants. When we asked participants to think about each step independently, they were able to articulate their rationale. However, for scenarios presented in long paragraphs, participants can express the overall comfortableness but struggle to articulate specific reasons.

Resulting design choice: We decided to break each scenario into multiple snippets since it allows participants to dive into the procedure of data practice and tell practitioners which part of the process might be problematic.

Risk–benefit analysis. Participants had difficulty in balancing the risks and benefits to make a confident decision/judgment. Throughout the interview, we asked follow-up questions to test participants' reactions to a few data practice variations. For example, besides personalizing the advertisements, the retail company can also use users' purchase behavior to optimize their supply chains to reduce the operation cost. We then asked participants whether knowing these benign data usages change their prior answers. In this process, most participants cannot consistently evaluate the potential benefits and harm, and elaborate their rationals as clear as the prior answer.

Resulting design choice: We decided to divide the risk–benefit analysis task into a set of single outcome evaluation tasks, where each participant only evaluates the impact of one outcome (i.e., one root-to-leaf branch in Figure 4) at a time.

4.3 LPR Story Representation

These findings motivated us to consider a new storytelling technique that describes a data practice from a data-centric perspective. An ideal privacy storytelling technique should be easy to understand for non-specialists and easy to generate for data practitioners.

Term definitions. We begin by defining four key concepts (Figure 4): data action, data application, data practice, and data stakeholders. We derive these terms from the past privacy discussions [56, 93, 115].

A *data action* is the smallest unit in the LPR story, which describes a specific operation where consumer businesses interact with users' data.[8]

[8]We adapted the definition of "data action" from [93].

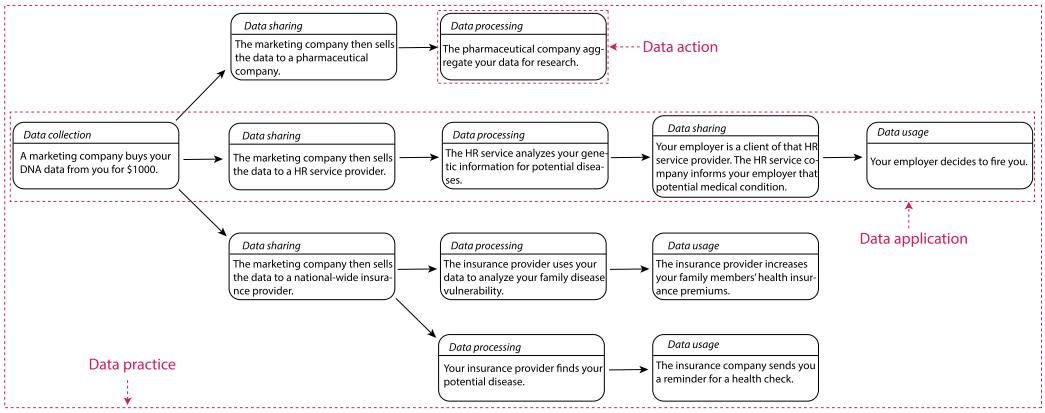


Fig. 4. A privacy storyboard of "trading your DNA data for $1000" [135]. Practitioners can transform an initial data practice design into a privacy storyboard using LPR's data action worksheet. We annotate the examples of "data action," "data application," and "data practice" in red boxes. In the storyboard, each node is associated with a data action. This tree topology explains the data flow using four data action vocabularies: collection, sharing, processing, and usage. This modular design allows practitioners to modify and append more data applications/actions incrementally.

A *data application* is a chain of data actions, describing the complete data lifecycle resulting in a specific consequence. For example, Google's 41 shades of blue data science experiment can be described as a chain of three data actions (Figure 2, left): a company first collects clickthrough data in an A/B test environment, then aggregates clickthrough data to find the one specific shade that generates most clicks, and finally decides to change the link color accordingly.

A *data practice* is a set of relevant data applications. For example, the clickthrough data can also be used for other data applications, such as search result evaluation [50] and advertisement billing [106].

A *data stakeholder* is an individual or group that could affect or be affected by the data practice [118]. We articulate three types of data stakeholders: data subjects, data observers, and data beneficiaries/victims. Data subjects are the people who contribute their data [56]. Data observers are the entities (e.g., people, algorithms, and companies) that have access to users' data [56]. Data beneficiaries/victims are the people impacted by a data practice.

Tree[9] topology. A common way for practitioners to present the information flow within a system is with data flow diagrams [77, 114], which visualizes the data communicated between different system components and external entities. For example, a data flow diagram of an online bookstore may describe how the shopping cart module sends the transaction data to the database. Such a system-centric graph can help the developers understand the system requirements, but do not necessarily explain data flow from the users' perspective.

In response to these limitations, we change the depiction perspective to a user-centric model, which describes how data flows between stakeholders and how data impacts different stakeholders. In doing so, we propose new data flow vocabularies to help practitioners communicate complex real-world privacy matters. While each node in a traditional data flow diagram is often a system component, the basic vocabularies in LPR are individual data actions, each documenting the privacy-relevant information, such as what data is being collected, where it is being sent to,

[9]The tree-like topology refers to the term in computer networks. Each node in a tree has zero or more child nodes but at most one parent node.

and why. Such a perspective shift allows non-specialists to track the accountability of different stakeholders explicitly.

A privacy storyboard then organizes these data actions in a tree topology. Figure 4 illustrates an example of an LPR story, which describes the data practice "trading your DNA data for $1000" from [135]. Each node in the tree topology is associated with a data action, and the edges connecting nodes represent the data flow between different data actions. Each root-to-leaf chain describes the data lifecycle of a specific data application. Such a tree topology makes the storyboard modular, so practitioners can incrementally modify and append more data applications/actions to the data practice.

Data action primitives. Through iterative development, we noticed that the data actions are clusterable. Our catalog of data actions is inspired by Solove's taxonomy [115] of harmful privacy activities, where Solove enumerates 16 activities that invade privacy and organizes them into four sequential groups: (1) information collection; (2) information processing; (3) information dissemination; and (4) invasion. While Solove originally uses these groups to organize harmful activities, we find that these groups are succinct and expressive vocabularies to describe the data flow.

We first extracted 67 data applications from 14 privacy scenarios[10] in Woodruff et al. [135] as the test cases. We then coded these applications into tree typologies and iteratively refined the definition of data action primitives. We eventually concluded with four types of data action primitives (Table 2): (1) Data collection; (2) Data sharing; (3) Data processing; and (4) Data usage. We also made following two main changes to make these vocabularies expressive in describing a data practice:

- **Data actions are neutral.** The original terms from Solove's taxonomy refer to problematic privacy activities. In contrast, we design LPR to evaluate an unknown data practice (malicious or benign), so we make data action primitives neutral and extend the coverage. For example, we change "invasion" to "data usage," which refers to the potentially neutral application a business can perform through the use of users' data that will impact the users in a certain way.
- **The order of data actions is flexible.** While Solove's order[11] describes the most common sequence of privacy invasions, we made it flexible to accommodate various complex real-world privacy-related data practices (see examples in Appendix Figures 23–34). For example, an HR service provider may purchase the data from a marketing company and then analyze the data (Figure 4).

4.4 Generating an LPR Story

Through an iterative coding process, we developed a method to help data practitioners think through a data practice and generate an LPR story. Instruction worksheets are presented in Appendix B. Here, we describe the three major steps and explain the design rationale.

(1) Identify data applications

- *Who are the data subjects? Who are the data observers? Who are the beneficiaries and victims?*
- *How would the data practice impact the stakeholders directly or indirectly?*

(2) Break each data application into data actions

- *Each data action should only contain at most one stakeholder per type (i.e., data subjects, data observers (senders or receivers), and beneficiaries/victims).*

[10]The appendix of [135] contains 20 scenarios. We excluded the non-consumer business ones.
[11]collection->processing->dissemination->invasion.

Table 2. A Data Action Is the Smallest Unit in LPR Story, which Describes a Specific Operation that Consumer Businesses Interact with Users' Data

Primitives	Definitions	Examples
Data collection	A data observer collects/stores data from data subjects.	Collection and storage
Data processing	A data observer processes users' data to derive new data.	Re-identifying, anonymization, inference, and aggregation
Data sharing	A data observer shares user's data or derived data with another data stakeholder (i.e., observer, subject, and beneficiary/victim).	Sharing, dissemination, transfer, and increasing data accessibility
Data usage	A data observer uses the data in a certain way that impacts a data beneficiary/victim.	Decision interference and knowledge discovery

We have four types of data action primitives: (1) Data collection; (2) Data sharing; (3) Data processing; and (4) Data usage.

(3) Describe each data action in succinct text

— *What is the context of the data action?*

Below, we examine each of these steps in detail, describing what kinds of information the question is looking for, why it is important, and offering some examples.

4.4.1 Identifying Data Applications. The first step for data practitioners is to enumerate the potential outcomes and applications that will be generated by the data practice in question. Data collected to achieve beneficial objects may adversely affect individuals' privacy as an unintended consequence [16]. For example, search engines use users' locations to make search results more relevant; however, the data collected might also be used to improve targeted advertising. Our goal here is to separate the duties [132] of different data applications by disseminating the tasks and associated privileges.

It would be challenging to enumerate every possible outcome; the focus here should be on most possible and common cases (both positive and negative). As such, we adapted two sets of questions from the privacy risk model by Hong et al. [56]. The first set asks data practitioners to think about the stakeholders impacted by the data practice: *Who are the data subjects? Who are the data observers? Who are the beneficiaries and victims?* The second set asks the author to think about the likely outcomes for each data application: *How would the data practice impact the stakeholders directly or indirectly?*

4.4.2 Breaking Each Data Application into Data Actions. The second step is to segment data applications into data actions and then organize them in a chain format. Our earlier pilot study (Section 4.1) suggests that the breakdown can both help participants more effectively understand the data practice and help practitioners to pinpoint specific problematic data action. The challenge here is to determine the segmentation strategy. An ideal strategy should produce relatively consistent segmentation results, which are easy to understand for non-specialists, while also be easy to use for practitioners.

The design of our segmentation strategy is inspired by the literature in UX storyboarding [121] and comic creation [79]. For example, Truong et al. [121] suggest that the optimal length of a storyboard is between three and five. Both undersegmentation and oversegmentation will decrease user understanding and engagement. We empirically tested different segmentation strategies suggested by McCloud [79], such as subject-to-subject, scene-to-scene, action-to-action, and moment-to-moment, iteratively on 94 data applications (67 from [135] and 27 from Table 1).

We finally concluded with the following segmentation strategy: each data action should only contain at most one stakeholder per type (i.e., data subjects, data observers, and beneficiaries/victims). For example, if a data action involves two beneficiaries, this data action should be further divided into two data actions. In doing so, the crowd workers can express their opinions to stakeholders separately. Meanwhile, the story authors can more easily trace accountability by tracking the data flow. Take the Target pregnancy prediction scenario [53] as an example. We can divide the data application into three data actions: (1) a retail company collects users' purchase data from their customers; (2) the retail company processes users' purchase history to predict if a female customer is likely to be pregnant; and (3) the retail company sends the tailored coupons to these mothers-to-be.

Note the final segmentation output still depends on the granularity of stakeholders. For example, practitioners may treat the whole retail company as a stakeholder so that they can test users' general perception to different prediction types (e.g., first-time parents and sports fans). Alternatively, they can also treat each department (e.g., data science department and marketing department) as an independent stakeholder so that they can trace accountability.

Once segmented, data applications may share common data actions. For example, the data collection node is often the root node of a tree, deriving multiple data usages. Merging the shared data actions results in a tree topology.

4.4.3 Describe Each Data Action in Succinct Text. The last step is to generate a succinct text description for each data action, explaining the privacy-salient information to a non-tech-savvy audience (e.g., crowd workers). There exist several description templates in different privacy frameworks, e.g., the five-parameter information flow template [6, 92] from the Contextual Integrity framework, and the examples in the PIA template [36]. We applied these templates to over 100 data actions and found two challenges in using them directly in the context of LPR.

- **Existing templates primarily focus on the data collection part but barely apply to other data actions.** Taking the Target pregnancy prediction scenario [53] as an example again, the five-parameter information flow can apply to the first data action (i.e., data collection): a retail company (recipient) collects their customers' purchase data (attribute) if they (sender) make purchase in online/offline stores (transmission principle). However, this template barely applies to the other two data actions (i.e., processing and usages). Data processing actions infer some new data insights from the raw data, data usage actions involves data beneficiary/victim. The five-parameter template does not cover both new data insights and beneficiaries/victims.
- **Existing templates are often too compact to include important contextual information.** In the Target pregnancy prediction scenario, the underlying goal of sending first-time parents special coupons is that the company wants to re-shape their life-long shopping habits. Research shows that first-time parents are experiencing the most hectic time in their life, and their shopping habits may experience a major change during that period. It is hard to fit this kind of contextual information into existing compact templates.

We then iterated text descriptions for the 94 data applications, tested them with crowd workers, and concluded with a lightweight description structure: [Context] [Data action][Further description]. Each data action description starts with an introduction of the context, then presents the specific data action, and concludes with a few necessary further descriptions. Reviewing the final descriptions, we propose a set of questions (attached in Appendix B) intended to help practitioners think through each data action and generate the descriptions. Table 3 enumerates some example output after this step.

Table 3. We Can Describe the Primary Data Application Derived from "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did" [53] as a Chain of Three Data Actions (Collection–Processing–Usage)

Data collection	When a user made purchases at a retail store (online/offline), the retailing company collected various behavior data, such as the purchase items, payment credit card, e-mail address, delivery address, and so on.
Data processing	Based on the e-mail, delivery address, credit card number, and so on, the company associates the purchase history with anonymized identity. The retail company later develops various models to predict user traits using the purchase history. For example, pregnant women purchase different items such as supplements like calcium, magnesium, and zinc. So the model guesses the customer may be pregnant if she buys these products in a short period.
Data usage	Once the retail company predicts the customers might be pregnant, the company will send out tailored coupons to these parents-to-be, e.g., special coupons for diapers, baby clothes catalogs. Sending coupons for baby items does not bring too much profit. The underlying goal is that the company want to re-shape their shopping habits. Research shows that first-time parents are experiencing the most hectic time in their life, and their shopping habits may experience a major change during that period. The retailing company wants to leverage this special opportunity.

The above table illustrates the example output by authors using the worksheet.

5 SCAFFOLDING THE PRIVACY DESIGN REVIEW

This section describes the design and development of a non-specialist-oriented privacy inspection, which collects users' qualitative (i.e., free-text responses) and quantitative (i.e., numerical comfortableness scores) feedback regarding their privacy concerns towards each data action (see Figure 2, middle).

5.1 A Failed Attempt: Heuristics for Crowd Privacy Inspection

In our initial task iterations (see Section 3.3), we attempted to develop a set of heuristics to capture common privacy concerns. Similar to UX heuristic evaluation [91], we presented participants with a set of heuristics and asked them to inspect each data action using these heuristics. In doing so, we hoped to be able to associate open-ended responses with different heuristics and present an aggregatable view to data practitioners.

We experimented three sets of heuristics for privacy problems (Table 4) derived from a range of legal and policy documents [1, 35, 43] and tested each set with crowd workers. For each set, we tested at least 5 different privacy scenarios with 10 crowd workers. Figure 5 illustrates one set of our heuristic questions. Participants were asked to go through the heuristics one by one for each data action, evaluate whether the data action violates them, and explain the reasons. In each round, we analyzed the collected responses to inform the task design (i.e., question templates and heuristics) in the next round. We observed a few tradeoffs through the process.

+ The heuristics can increase crowd workers' sensitivity to privacy problems. With heuristics, participants become more critical and verbal about their privacy concerns, tending to mention one or multiple privacy problems for each heuristic.
+ We can use heuristics to aggregate users' open-ended responses.
– Participants used the heuristics as the criteria rather than their actual feelings. For example, one participant reported that "sending personalized coupons" violates "respect for persons," but also noted that he is "fine with it since it may save him some money."

Table 4. Sources We Used to Generate Heuristics for Privacy Problems

#	Source	Heuristics
1	Seven types of privacy [43]	Privacy of the Person, Privacy of Personal Behaviour and Action, Privacy of Communications, Privacy of Data and Image, Privacy of thoughts and feelings, Privacy of location and space, and Privacy of association
2	Privacy rights by Roger Clarke [1]	Privacy of the Person, Privacy of Personal Behaviour, Privacy of Personal Communications, and Privacy of Personal Data
3	Belmont Report [35] (Figure 5)	Respect for persons, Beneficence, and Justice

Fig. 5. One survey screenshot of the failed attempts using heuristic guidelines. Participants were asked to go through the heuristic one by one for each data action, evaluate whether the data action violates it, and explain the rationale. The example heuristic guidelines are derived from Belmont Report [35]. While heuristics makes crowd workers more sensitive to privacy concerns, they also introduced two tradeoffs: (1) participants become overly critical; and (2) participants either have a shallow understanding of the heuristics or spend significant time to learn the heuristics.

- Participants tend to check all the minor privacy problems and have a hard time prioritizing them. As a result, the collective output of the heuristic-based approach quantifies both how visible the problem is, and how severe the problem is.
- Learning the heuristics during a short survey was challenging. Most crowd workers are not aware of the principles before joining the tasks. They can only gain a limited understanding during the study instruction and warm-up tasks.
- Going through the heuristics one by one was time-consuming. Inspecting four sequential data actions with seven heuristics often takes more than an hour. Meanwhile, the response quality from crowd workers is often reduced as the completion time increases. In the first four iterations, we aggressively reduced the number of heuristics to speed up the task. However, the last heuristic-based approach using three heuristics still takes around 45 minutes to complete a survey.

These findings motivated us to rethink the design of the privacy review task and the necessity of heuristics. Privacy review by crowd workers is different from heuristic evaluation in several respects. First, heuristic evaluation is mainly conducted by trained UX researchers since it takes time and practice to gain a deep understanding of usability heuristics. Second, privacy opinions

can be more personal and abstract than usability problems and therefore can be difficult for users to pinpoint what's bothering them and map it back to heuristics.

In the later iterations, we shifted the design of privacy review tasks to a two-stage design, which breaks the task into two sub-tasks: inspection (Section 5) and annotation (Section 6). We first asked participants to provide free-text responses without heuristics, and then forwarded these responses to a different set of crowd workers to annotate privacy concerns. This inspection–annotation paradigm has several advantages over the heuristic-based approach. First, users tend to describe the privacy concerns they care about most in free text, which helps us prioritize their privacy concerns. Second, users do not need to learn and reflect on the heuristics, so it reduces the heuristic overhead and speeds up task completion. Meanwhile, we can aggregate users' open-ended responses using the labels from the annotation tasks.

In the following subsections, we introduce three important features of LPR, which further scaffold the privacy inspection task for crowd workers: (1) eliciting users' feedback through privacy concerns (Section 5.2); (2) searching privacy problems collectively and actively to increase the coverage (Section 5.3); and (3) dividing the survey into mini-surveys to lower the task barrier (Section 5.4). We present the design of the annotation tasks in Section 6.

5.2 Eliciting Users' Feedback through Privacy Concerns

These early explorations also raised an important question: what do we aim to measure through LPR (e.g., privacy problems, privacy concerns, and privacy risks), and how do we guide participants to inspect the target properties?

Indeed, privacy is a concept in disarray [115]. Researchers and practitioners have developed many terms to capture different dimensions of this complicated concept. For example, Solove [115] articulates the term "harmful privacy activities" by enumerating 16 activities that invade people's privacy. Privacy engineering techniques (e.g., PIA [136] and PRAM [93]), on the other hand, use the term "privacy risk," which is the multiplication of the likelihood and the impact of privacy harms. We surveyed relevant privacy concepts in the literature (Table 5), tested multiple privacy concepts in experimental surveys, analyzed the responses from crowd workers.

We chose to focus on the concept of privacy concerns. The term "privacy concern" has been heavily used by privacy researchers in the HCI community [42, 67, 101, 112], which refers to users' subjective feeling (e.g., whether users are comfortable) towards a specific privacy relevant problem. Different from privacy risk, privacy concern is not necessarily associated with personally identifiable information collection or potential privacy harm. For example, users might feel uncomfortable because of an unfair procedure, a digital market manipulation [17], or simply a lack of trust.

We decided to focus on privacy concerns for two reasons. First, users' subjective feelings towards a data practice are important, yet there exists few systematic methods to capture them directly at a low cost. Second, both our formative studies and early privacy surveys also found that users have a desire to express their privacy opinions. It is easier to ask users to reflect on their feelings rather than evaluate the data practice using other system metrics (e.g., privacy risks).

5.3 Searching Privacy Problems Collectively and Actively

Existing research often formulate privacy concerns as numerical values and measure them using vignette factorial surveys [42, 110]. For example, Schechter et al. [110] uses the question "If someone you cared about were a candidate participant for this experiment, would you want that person to be included as a participant?" to collect respondents' responses. Respondents can answer in three options: "Yes," "I have no preference," or "No." To measure users' concerns

Table 5. A List of Commonly Used Privacy Concepts

#	Privacy concept	Description	References
1	Privacy harm/injury/ activity	the negative impact of a data action, which can be physical, mental, or economic injury to the consumers	[24, 93, 115]
2	Privacy risk	the likelihood of a privacy harm multiply the impact of the privacy harm	[56, 94]
3	Perceived privacy risk	users' perceived likelihood of a privacy harm multiply users' perceived impact of the privacy harm	[12, 44, 86]
4	Privacy concern	a combination of perceived privacy risk/benefit tradeoff, procedural fairness, trust, and feelings	[42, 67, 112]
5	Privacy norm/ expectation	established based on the collective pattern of privacy concerns	[6]
6	Privacy preference	established based on an individual's consistent privacy concerns on similar data practices	[73]
7	Privacy violation	a problematic data action that would violate privacy norms	[102]

We articulate these privacy concepts to illustrate the conceptual differences.

Table 6. Comparison between LPR and Existing Techniques: Privacy Engineering Techniques (e.g., PRAM [136] and PIA [136]) and FVS

Techniques	Who discovers privacy problems?			Who quantifies privacy problems?		
	Experts	Practitioners	Laypersons	Experts	Practitioners	Laypersons
PRAM [93]	✓	✓		✓		
FVS [110]	✓			✓		✓
LPR			✓			✓

Methods like PRAM, PIA require privacy engineers to discover and quantify potential privacy problems. FVS requires researchers to enumerate privacy problems and quantify the issues by analyzing laypersons' responses. In LPR, laypersons both discover and quantify privacy problems.

from these responses, researchers need to survey a large number of users [110], carefully control the variables across different surveys, and quantify users' concerns through statistical analysis (Table 6). This process is often expensive and time-consuming. Meanwhile, practitioners may not have the necessary skills to run such a rigorous study.

In contrast, LPR contributes a different and complementary method. Our method is inspired by heuristic evaluation [84, 91]—we make an analogy between usability problems and privacy problems. In traditional user testing (Table 7), the observer (i.e., experimenter) has the responsibility of interpreting users' actions to infer how these actions are related to the usability issues. Heuristic evaluation reduces the cost of finding usability problems by asking evaluators to actively assess the usability of a user interface using a set of heuristics.

The conventional approach of using FVS (Table 8) operates similarly to traditional user testing, which requires the experimenter to interpret respondents' numerical answers. Instead, LPR asks users to nominate their privacy concerns actively, which is similar in spirit to heuristic evaluation. In doing so, LPR formulates privacy concerns as *the reasons why users feel uncomfortable regarding specific data action*. The input from the crowd workers would be a set of free-text responses associated with a severity score, each explaining why users feel uncomfortable regarding a specific data action.

Table 7. Usability Evaluation [90]

Traditional user testing	The experimenter interprets users' action to infer how these actions are related to the usability issues.
Heuristic evaluation	The evaluators go through the interface from users' perspective, assess the usability and report their comments.

Table 8. Privacy Concern Inspection

Privacy factorial surveys	The experimenter designs surveys consisting of varying situations and interprets the respondents judgments.
LPR	Crowd workers go through a data practice from users' perspective, assess the privacy concerns and report their comments.

In practice, LPR asks participants to answer two questions (see an example in Figure 2, middle) regarding each data action independently[12]:

— *How would you feel if the company [collects|shares|processes|uses] your data as described above?*
This is a five-scale likert-score question. Users can answer in five options (from "Extremely comfortable" to "Extremely uncomfortable").

— *Why did you feel that way?*
The is a open-ended question, which requires 100-character [3, 120] minimum length. Participants are asked to write the response in a template: "I feel [comfortable| ... |uncomfortable] because XXXXX."

Through our survey iterations, we also found our participants were quite expressive in describing their major privacy concerns. The reported concerns also varied across different participants, so collectively aggregating them helped improve coverage. We quantitatively evaluate the benefits of aggregating inspection results in Section 9.

5.4 Dividing the Survey into Mini-Surveys

The above design allows a non-tech-savvy participant to contribute to a privacy review. However, the length of the reviewing task still remains challenging for crowd workers. Inspecting a complex data practice can be time-consuming, taking up to two hours in our early surveys, which is a long period of time for many crowd workers [108]. On the other hand, arbitrarily breaking down a data practice inspection into smaller parts might not work either since data actions are highly interdependent. Breaking a data practice inspection into data action inspections might remove useful contextual information, making it hard for participants to understand the necessary privacy details. After testing three "divide and conquer" strategies (dividing up the survey by actions, by applications, and by practices), we decided to ask users to inspect the data practice at the granularity of data applications.

Each crowd task (i.e., **Human Intelligence Task** (**HIT**)) contains four parts: consent page, tutorial examples, data application evaluation, and demographic data collection. In the tutorial example page, we showed participants the example answers for a data application derived from [135]. We used a Depth-first search algorithm to parse the tree typology and generate a set of data applications. The LPR-web then loads a pre-defined survey template and replaces the data action description placeholders (see Figure 13) with the actual content in each data application. Each crowd task contains one to two data applications, and each survey page renders the evaluation questions for one data action. It often takes around 15 minutes for a crowd worker to complete a survey. A participant is only allowed to work in one data practice once.

[12] Adapted from [42].

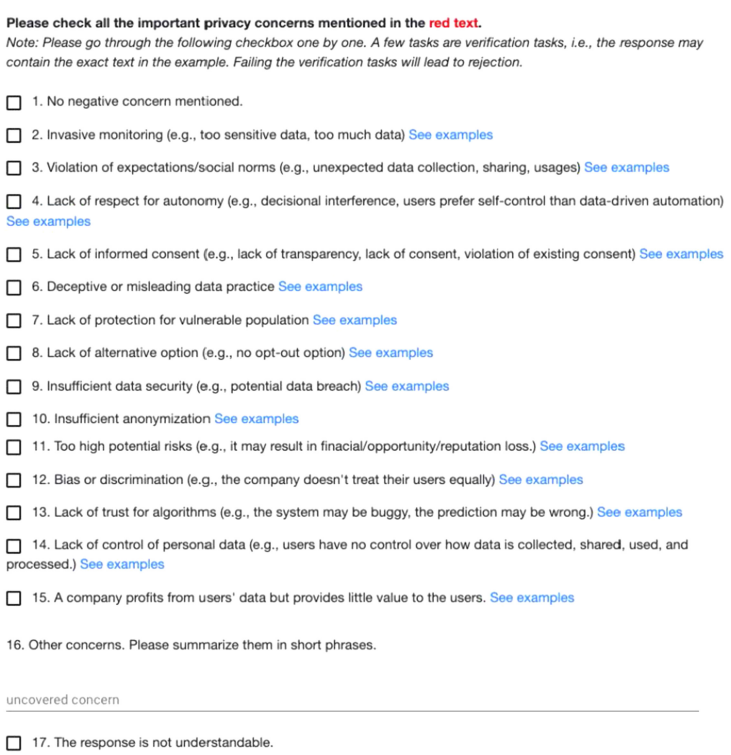


Fig. 6. The final interface of LPR privacy annotation task. We ask each crowd worker to go through the privacy concern categories one by one and check all the concerns mentioned in users' free-text responses (collected from Figure 2(b)).

6 AGGREGATING THE PRIVACY INSPECTION RESULTS

A major challenge of introducing open-ended questions into privacy surveys is that they can be difficult to aggregate. For example, Felt et al. [42] used open-ended questions in their surveys and manually coded these responses to interpret users' contexts. However, practitioners may not have the qualitative research skills or time to analyze these open-ended responses. To address this challenge, we developed a crowd-based technique to aggregate open-ended privacy opinions, which transforms the free-text responses and the comfortableness scores into a list of privacy concerns, each associating with a magnitude score.

6.1 Privacy Annotation

We organized the annotation tasks in a manner similar to traditional image labeling tasks. Each free-text response annotation is an independent HIT, and participants can complete as many tasks as they desire. We then presented each participant the data action description, the response text, and a set of privacy concern options. The annotator can select multiple privacy concerns or nominate new privacy concerns not covered in the provided list. In our experiments, we assign each response text to three different participants and consider an annotation valid if more than two participants annotate the same concern independently.

A list of common privacy concerns. At the core of the annotation is the list of common privacy concerns (Appendix A), which is the basic vocabulary to characterize users' privacy concerns in LPR, facilitating communication between crowd workers and practitioners. Designing

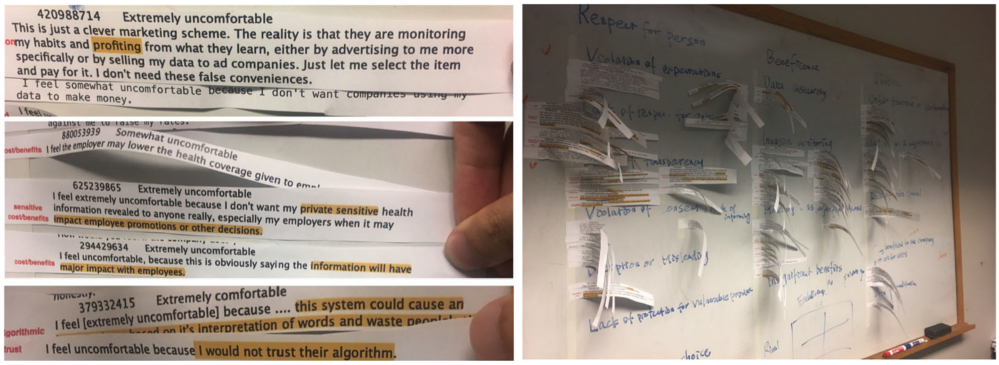


Fig. 7. The development of a list of common privacy concerns. We first applied thematic analysis to our survey and interviews data to build the initial list. The left figure illustrates some example line-by-line coding results based on the survey responses. We then iterated the list with crowd workers, to ensure the list is easy to comprehend for non-specialists. We placed the example quotes, basic codes, and high-level themes on a whiteboard to track the development (right).

such a list is challenging. First, the list should be easy to understand for crowd workers, imposing little learning overhead. Second, the list should be comprehensive, covering most privacy concerns people may have. Third, the granularity of the list should be meaningful for practitioners, ideally connecting to some actionable items.

We used thematic analysis [14] to build the initial list. Two authors read through the free-text responses collected in survey iterations as well as the notes from previous semi-structured interviews independently and held weekly discussion. The aim of the analysis is to find out privacy concerns that the study participants had about our data practices. With this idea in mind, we performed the following steps of analysis: line-by-line coding of the survey and interviews' transcripts, with an aim to generate as many basic codes as possible, constant comparison and discussion of these basic codes, development and refine of higher-order themes to encompass and link these basic codes, and extensive case-based memo writing to track developing ideas. As the analysis proceeded, we also developed new privacy storyboards, refined the survey templates, and collected responses from different sets of crowd workers. We iteratively analyzed the data we collected throughout the process until we found consistent, recurring types of privacy concerns and no longer encountered new types of concerns, which suggested that we had reached data saturation. In total, we analyzed over 1,000 responses to 67 different data applications.

We then iterated the list of privacy concerns and description text with crowd workers, to ensure the list is easy to comprehend for non-specialists. We further optimized the list for annotation efficiency by merging several conceptually relevant concerns, as we found that crowd workers often labeled them together. To help practitioners comprehend the list quickly, we further organized these privacy concern types into three high-level categories: respect for persons, beneficence, and justice (see Appendix Figure 22). Figure 6 enumerates the text descriptions of the 14 concerns. Note the list of privacy concerns is non-exclusive, and so a free-text response may fall into multiple categories.

6.2 Computing the Range and the Magnitude of Different Privacy Concerns

Through the privacy inspection surveys, a practitioner obtains a set of free-text responses $\{t_1, t_2, t_3 \ldots\}$ associated with severity scores $\{s_1, s_2, s_3, \ldots\}$, each explaining why users feel uncomfortable/comfortable regarding a specific data action. She then obtains a set of privacy

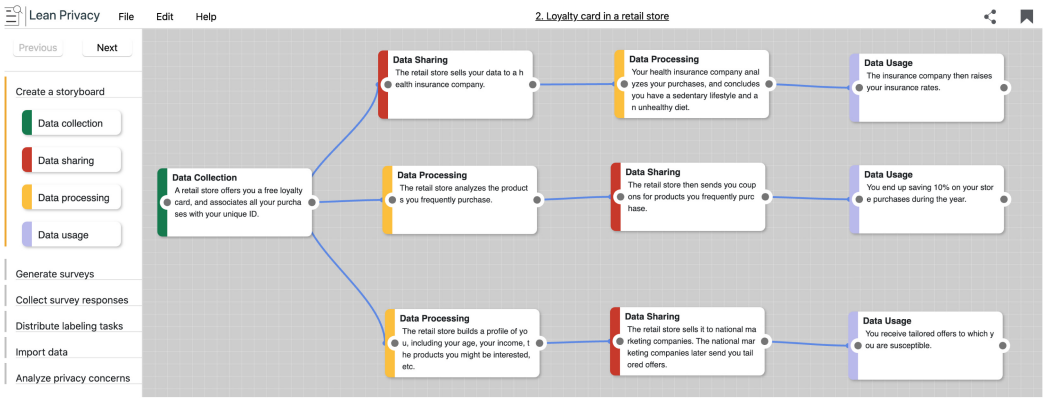


Fig. 8. The overview of an LPR story in the web interface. The tree topology makes the story modular, so practitioners can incrementally modify and append more data actions.

concern annotations $A_k \sim \{a_{k1}, a_{k2}, a_{k3}, \ldots\}$ for each free-text response $\{t_k\}$ through the annotation tasks.

We define the range of privacy concerns of a data action as the union of annotated privacy concerns for individual free-text responses: $Range = A_1 \cup A_2 \cup A_3 \cup \ldots$. The range of privacy concerns of a data practice is the union of individual data actions' privacy concerns. We then quantify the magnitude of each privacy concern similarly as privacy risk estimation [93]. In essence, the magnitude is a function of the likelihood that a privacy concern expressed by the crowd workers multiplied by the level of their uncomfortableness. LPR quantifies the magnitude of each privacy concern as the sum of associated comfortableness scores from different participants divided by the total number of participants.

7 A WEB-BASED SYSTEM TO STREAMLINE LPR

We further developed a web-based system[13] to streamline the LPR process. A practitioner may use the worksheet (Appendix B) to think through the data practice in the lens of LPR data actions and then create a storyboard using the web interface (Figure 8). The practitioner can continuously modify the stories, append new applications, and make changes to the text descriptions. All the modifications would be saved to a remote server automatically.

Once finished, the practitioner may download a **JavaScript Object Notation (JSON)** file that can be imported into the Qualtrics survey platform.[14] The practitioner can then distribute the generated surveys (Figure 2) through various channels (e.g., **Amazon Mechanical Turk (AMT)** and Google Marketing Platform), collect users' direct feedback, and import the survey responses to an HIT annotation template (Figure 6). At the end, the practitioner can import the **comma-separated values (CSV)** files into the web system to view them quantitatively. The web system displays the results in an interface inspired by Heuristic Evaluation [91]. It has two views to support practitioners to navigate and consume the crowd inspection output: summary view and detailed view. A complete demonstration of the web interface is available in the accompanying video.

Summary view. Once the crowd inspection results are available, the web system shows an aggregated privacy concern spectrum (Figure 9), which can help practitioners both gain a high-level overview and navigate to the specific comment by hovering mouse on a specific block.

[13] http://leanprivacyreview.com/.
[14] https://www.qualtrics.com.

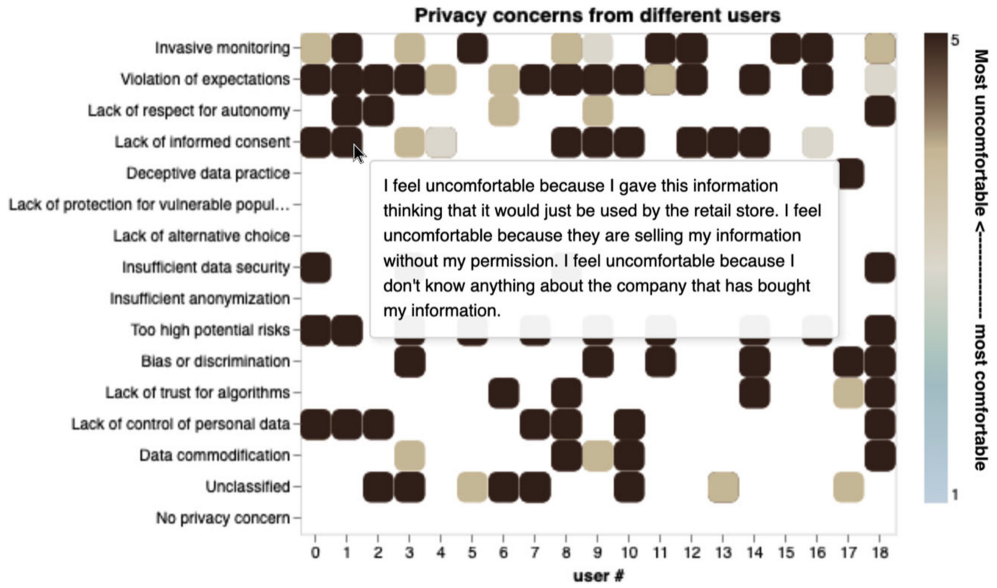


Fig. 9. The summary view of an example data practice: Loyalty card in a retail store [135]. The corresponding storyboard can be found in Figure 12. The privacy concern spectrum above renders privacy review results from 19 participants, showing who found which privacy concern. Each column corresponds to one participant, and each row corresponds to one type of privacy concerns. The redness in each block indicates the level of each participant's discomfort regarding the specific dimension. We do not include positive feedback in this privacy concern spectrum. When a practitioner hovers her mouse on a colored block, a tooltip containing the raw textual feedback will pop up. This grid plot is inspired in part by heuristic evaluation [91].

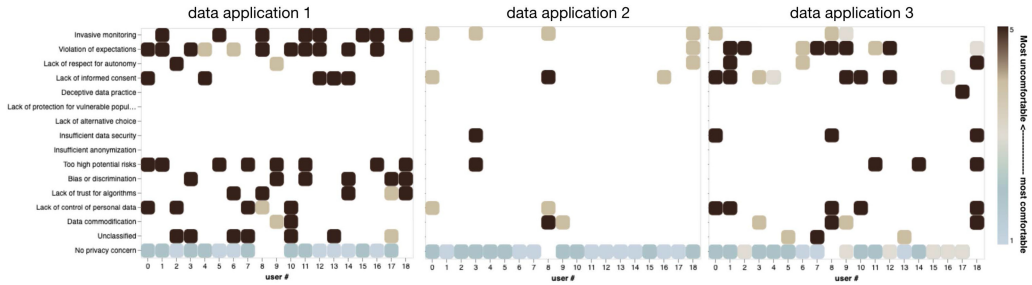


Fig. 10. The summary views at the data application level for Loyalty card in a retail store [135]. We only annotate the free-text responses associated with non-positive scores, so all the positive responses fall into the "no privacy concern" category. While the total number of free-text responses are the same, the summary views are quite different. At a glance, the practitioners can find that the data application 1 (using the purchase data to determine insurance rates) raises most privacy concerns, while the data application 2 (using the purchase data to customize coupons) receives least negative comments.

The usability problem visualization in Heuristic Evaluation [91] inspires the design of the spectrum, where each column corresponds to one participant, and each row corresponds to one type of privacy concerns. The redness in each block indicates the level of each participant's discomfort regarding the specific dimension.

Practitioners can view the privacy spectrum at the level of a data practice or a data application. For example, Figure 10 illustrates the spectra of three data applications in a data practice. At

Fig. 11. When the practitioner taps a data action node, LPR provides the details about the selected data action. The detailed view first shows a comfortableness distribution chart (left), helping the practitioner understand the general acceptableness. The detailed view then shows a concern bar chart (middle), visualizing the occurrences of privacy concerns in each category. Finally, the practitioner can explore the free-text responses by selecting the concern type of interest in the concern bar chart.

a glance, practitioners can find that crowd workers raise most privacy concerns in the data application 1 (Figure 10, left), while the least privacy concerns in data application 2 (Figure 10, middle).

Detailed view. When the practitioner taps a data action node in the review mode, LPR provides the details about the selected data action (Figure 11). The detailed view aims to help practitioners dive into the details and gain a deeper understanding of users' privacy concerns. The detailed view contains three components: (1) a comfortableness distribution chart that helps authors understand the general acceptableness (Figure 11, left); (2) a concern bar chart that visualizes the occurrences of privacy concerns in each category (Figure 11, middle); and (3) a filterable table that helps practitioners to read the raw responses associated with each privacy concern category (Figure 11, right).

8 CASE STUDIES ON REAL-WORLD DATA PRACTICES

We analyzed 12 real-world data practices derived from media reports and research papers using the storytelling framework outlined in Section 4. The purpose of these case studies is twofold. We use these case studies to test the expressiveness of the storytelling framework, and later used these generated storyboards for LPR evaluations (Section 9).

8.1 Method

We selected 12 data practices across a diverse set of categories, such as Internet of things (Appendix Figures 25 and 31), eCommerce (Appendix Figures 24, 25, and 34), social networks (Appendix Figures 28 and 30), advertising (Appendix Figure 23), computational psychology (Figures 26 and 33), data science experiments (Appendix Figures 23, 28, and 29), and scenarios involving vulnerable populations (Appendix Figures 27, 28, and 31). We intentionally excluded: (1) government surveillance since it often involves complex civil rights discussions; and (2) data breaches since these are not intentional parts of a system's design.

Individual reports or papers may only involve one data application; we combined the relevant stories if the underlying data actions are similar. For example, a social network company may use the emotion-sensing technique (Figure 28) to run a psychology experiment [65], a teenage suicide detection [27], or an emotion-based advertisement targeting [71, 104]. Here, we focused on the scenarios and outcomes that are most common and natural based on a review of media reports. Although we did not enumerate all possible cases, practitioners can easily extend the story to cover more data applications in the future.

Table 9. The List of 12 Tested Scenarios Derived from Media Reports and Research Papers

#	Scenario name	Description
1	Search engine click-through data [7, 55]	Google runs an A/B test experiment to determine the best color for the weblinks in search results.
2	Loyalty card in a retail store [135]	Woodruff et al. [135] enumerated several common data usages associated with retail store loyalty cards.
3	Checkout free retail store [81]	Amazon Go stores install hundreds of cameras to enable checkout-free shopping experience and other smart features.
4	Game chat log [100]	Riot uses League of Legends chatlogs to weed out toxic employees.
5	Pregnancy intimate data [51]	The pregnancy-tracking app Ovia lets women record their most sensitive data for themselves—and their boss.
6	Social network sentiment analysis [27, 65, 71, 104]	Facebook may use the emotion-sensing technique to run a psychology experiment [65], a teenage suicide detection [27], or an emotion-based advertisement targeting [71, 104].
7	Online dating experiments [54, 134]	OkCupid runs two data science experiments to understand the romantic relationship: score manipulation & love is blind.
8	Email contacts for social network bootstrapping [52, 88]	Google appropriates the G-mail data to bootstrap Google Buzz, a social networking service.
9	A fitness tracker [103]	Sexual activity tracked by fitbit shows up in Google search results.
10	Predicting a woman's pregnancy [53]	Target uses users' purchase data to predict a woman's pregnancy, aiming to hook parents-to-be at that crucial moment.
11	An insurer employs AI [96, 129]	China's largest insurer, Ping An, starts to employ artificial intelligence to identify untrustworthy and unprofitable customers.
12	eCommerce Price discrimination [2, 22]	(1) Travelocity practices price discrimination using users' device information. (2) Uber finds that users are more likely to pay for surge pricing if the battery is low on their phones.

We selected these data practices across a diverse set of categories, such as IoT (#3, #9), eCommerce (#2, #3, #12), social networks (#6, #8), advertising (#1), computational psychology (#4, #11), a data science experiment (#1, #6, #7), and scenarios involving vulnerable populations (#5, #6, #10).

Three participants were involved in the story creation process. These participants had backgrounds in User experience research, privacy research, and software engineering, respectively, mimicking a real world setting in the industry [28]. Two of the participants were not aware of the storytelling framework in advance. We first briefed participants on the framework, then presented them with a set of compiled news media reports, and finally asked them to summarize these new reports and generalize an LPR story representation collaboratively. Each story creation took between 20 and 30 minutes.

8.2 Results

Figure 12 illustrates an example privacy storyboard overview of data practice #2. We only offer a brief text summary for each data action in these illustrations. The complete set of all 12 illustrations is available in Appendix C.

9 EVALUATION: CONSISTENCY, COST AND LATENCY, AND RESULT QUALITY

We demonstrate through our evaluation that LPR is cheap, fast, consistent, and can help practitioners find the concerns they are not aware of. Note that our evaluation focuses on the

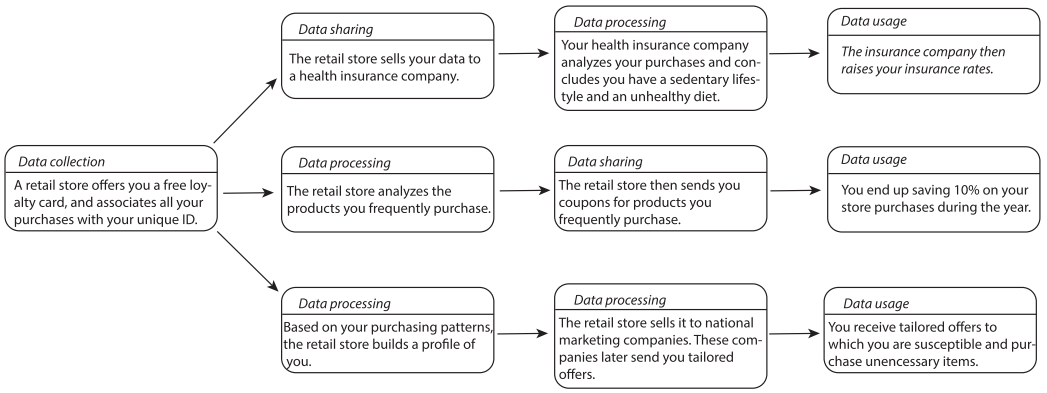


Fig. 12. A privacy storyboard of scenario #2 "Loyalty card in a retail store" using the LPR framework. A retail store collects users' data through a loyalty card and uses the data for insurance and coupon personalization. Three participants created the storyboard collaboratively using the worksheet (Appendix B).

feasibility of using crowd workers to learn about potential privacy concerns pertaining to data practices rather than the usability of our support tools (e.g., our website and our visualization). We leave these evaluations in future work.

Result summary: (1) Discovery of privacy concerns saturates as the number of evaluators exceeds 14 participants; (2) LPR can collect a consistent and saturated results in 5.5 hours, costing 3.7 hours of total crowd work (≈$80); and (3) The crowd inspection found 89% of privacy concerns identified by a group of software and data practitioners as well as 139% more privacy concerns that practitioners are not aware of, at a 6% estimated FA rate.

9.1 Deployment and Apparatus

We deployed the generated surveys (Figure 13, left) on AMT to test the practical applicability of LPR. We recruited 20 unique participants through TurkPrime [74] for each story and paid each participant $1 for completing a short survey (including less than 5 free text questions) or $2 for a longer survey. We also informed our participants that we will give an additional $1 reward if participants provide thoughtful reasoning in open-ended questions. To avoid priming workers during task selection [15], we advertised the survey as a "data practice survey," not as a survey specifically about privacy. On average, it takes 8 minutes to complete a short survey and around 15 minutes to complete a long survey. We granted bonus rewards to all the participants. So the average hourly pay for an inspection participant is between $12 and $15.

We then distributed the neutral[15] and negative free-text responses in privacy concern annotation tasks (see Figure 6) on AMT. Here, annotators are allowed to complete multiple tasks. We forwarded each free-text response to three annotators and rewarded each completed task $0.20. LPR accepts the annotation if more than two participants annotate the response into the same concern. Each annotation task takes around 45 seconds, resulting in an hourly wage of $15.

Baseline. Our baseline is to mimic the real-world process: a small team (N = 6) of practitioners need to guess users' reactions to justify the privacy design [20] when a formal privacy review is not available. We recruited 24 data practitioners (11 male, 13 female) through mailing lists and campus flyers. All participants have advanced degrees: 3 Bachelor, 14 Master, and 5 Ph.D. Participants also indicated that they have at least two years of experience in one of the following areas: User Experience Research (11), Software Engineering (7), and Privacy Engineering (6). A total of 16

[15]We also forwarded the neutral responses since some neural responses contain both positive and negative opinions.

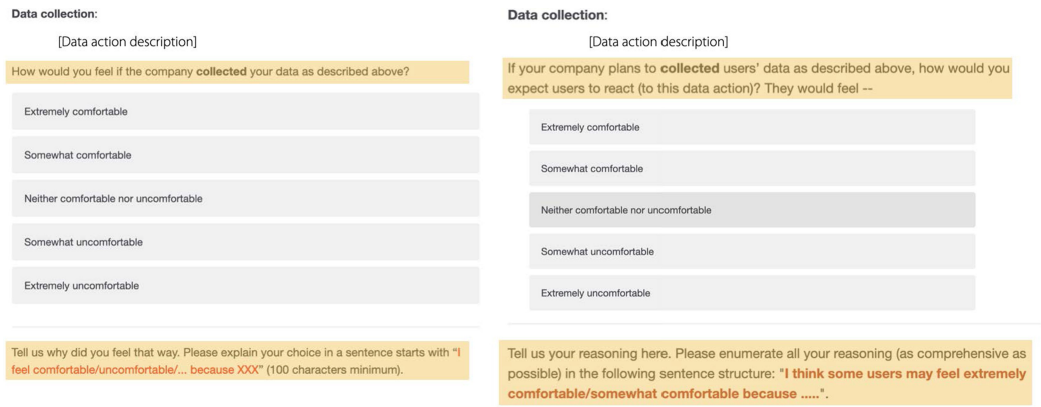


Fig. 13. The survey templates for crowd workers (left) and the software and data practitioners (right). The left template asks participants (i.e., crowd workers) to imagine themselves as users. In contrast, the right template asks participants to imagine themselves as company employees and guess users' reactions. We highlight the differences between the templates in yellow boxes.

out of 24 participants have professional working experience in an IT company. We consider these participants have basic-to-intermediate knowledge about privacy, who have a comparable level of privacy expertise as the real-world teams [28].

We then tailored the survey template for practitioners (Figure 13, right), in which we asked respondents to assume themselves as company employees, guess users' responses, and explain their rationale. We asked each data practitioner to review three different stories and complete the corresponding surveys, which took between 40 and 70 minutes. We compensated each participant $20 for their time.

Ground truth. We recruited three trained HCI privacy researchers (one fourth-year Ph.D. student and two Post-doc researchers, one male and two female) to create the ground truth list of privacy concerns for each data action. None of these researchers participated in the earlier story creation process nor the authors of this article. All three researchers have rich experience conducting user studies regarding users' privacy concerns and have worked on usable privacy research for more than four years. These experiences allow them to anticipate privacy concerns better than most data practitioners. Although individual experts may suffer from biases and blind spots, we aggregate all three researchers' judgments to leverage the collective wisdom.

We first presented these researchers with the privacy concern taxonomy (Appendix A) and briefly introduced the structure of the taxonomy. We then merged the results (i.e., lists of privacy concerns) from the crowd workers and the practitioners, and presented the combined results **blindly** to researchers. For each data action, researchers saw the text description, a list of merged privacy concerns, and the complete taxonomy. To avoid biases, researchers were not aware of the sources of individual privacy concerns.

We then asked the researchers to evaluate each privacy concern's importance on a 10-point scale and identify the "false positive" concerns (FAs). We defined importance by adapting the definition of privacy risks into the context of privacy concerns: the likelihood multiplies the severity. We considered a privacy concern a FA if more than two judges marked it invalid. The inter-rater reliability for labeling FAs was 0.82. We computed the ground truth list with magnitude by summing the ratings from individual researchers.

Caveats. The ground truth list may not cover some "unknown unknown" privacy concerns if all the participants (i.e., the crowd workers, practitioners, and researchers) miss these concerns.

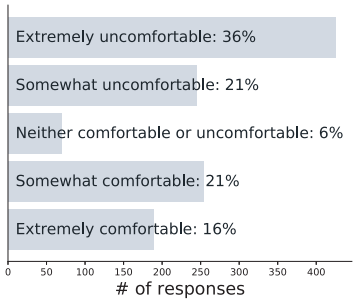


Fig. 14. The aggregated distribution of responses in all scenarios across different scores. We present the individual distributions in Appendix D.

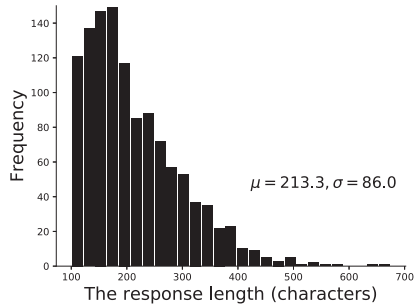


Fig. 15. The aggregated frequency distribution of different response lengths in all 12 scenarios (Appendix Table 9).

Each data practice in the experiment has been inspected by 20 crowd workers, 6 data practitioners, and 3 researchers. It is very unlikely that any of the data practices would have any major privacy concerns which did not bother some of these participants enough to complain about it. However, it is still possible that the next subject would stumble over a new minor privacy concern. Therefore, we stated the "known" number of privacy concerns for each story, and the statistics in Section 9.3.3 are based on the number of known problems.

9.2 Crowd Inspection Data Summary: Free-Text Responses and Annotations

In total, we collected 1,182 privacy opinions from crowd workers regarding each data action, each containing a short free-text response associated with a comfortableness score (five-point scale). Figure 14 illustrates the distribution of the responses across different scores. Most participants took a positive or negative position on the data action, while only 6% chose to be neutral (i.e., neither comfortable or uncomfortable). The average length of the free-text responses is 213.3 (std = 86.0) characters, and the longest response is 675 characters. Figure 15 illustrates the frequency distribution of different lengths.

We forwarded the neutral[16] and negative responses (#=739) in annotation tasks and collected three independent annotations for each response. Crowd workers annotated the aforementioned privacy concern using our provided list in 2,197 tasks and reported "uncovered privacy concerns" in the other 20 tasks (less than 1%). On average, each crowd worker checked 1.7 (std = 1.0) privacy concerns in a privacy concern annotation task (Figure 16). The average inter-rater reliability is 0.40 (std = 0.33), indicating that crowd workers agreed with each other moderately well. Since crowd workers often stop labeling privacy concern categories once they find one or two match concerns (avg = 1.7 out of 15), this agreement is lower than the agreement between privacy researchers.

For each uncovered privacy concern, participants provided a short description in the "other concern" textbox (Figure 6). We manually reviewed "other concern" annotations and found that they were often covered in our provided list. For example, one participant described a new concern—"involves invasion of women's right to privacy," which we felt fell into the category of "lack of protection for the vulnerable population." We design the annotation task as a lightweight task (30–60 seconds), so it is a tradeoff that some participants may have a shallow understanding of the list. Indeed, LPR can still identify an accurate annotation by merging the annotations from multiple participants since the "other concern" annotation is infrequent (less than 1%).

[16]We noticed that many neutral responses contain both positive feedback and negative concerns. Participants assessed their comfortableness as neutral after balancing both perspectives.

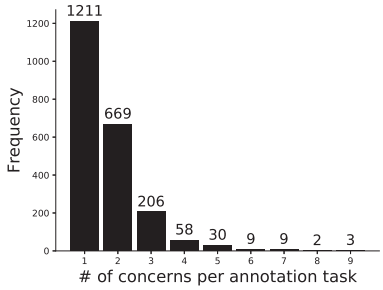


Fig. 16. The distribution of the number of concerns found per annotation task. Most annotation participants only check one privacy concern in a privacy concern task.

9.3 Results

The output of the crowd inspection is a list of privacy concern annotations regarding different data actions. Each annotation is associated with a short free-text response, a comfortableness score (five-point scale), and crowd workers' participation data (e.g., a start time and a completion time).

To characterize the LPR's performance, we summarize our results based on the following three criteria:

— Consistency: Would discovery of privacy concerns saturate as the number of evaluators increases? How many crowd workers does LPR need to achieve a decent coverage?
— Cost and latency: How much does the crowd inspection cost? How long do practitioners need to wait for the feedback?
— Result quality: How good is the range of discovered privacy concerns? Can LPR prioritize the concerns well?

9.3.1 Consistency. Our evaluation of the consistency is similar to the experiments Nielsen et al. [91] conducted for usability heuristic evaluation. We constructed hypothetical aggregates of varying sizes to test how many privacy concerns such aggregates would theoretically find. For each story, aggregates were formed by choosing the number of people in the aggregate randomly from the total set of crowd workers.

To recap, each privacy concern is always associated with a data action. If LPR finds that two different data actions (A, B) both have a privacy concern X, LPR recognizes A-X and B-X as two privacy concerns. However, if two free-text responses for the same data action mention a privacy concern X, LPR only considers it (i.e., A-X) as one privacy concern.

We further quantified the consistency in two settings: an explorative search and a tight-budget search. In the explorative search setting, we want to understand whether discovery of privacy concerns saturate as the number of inspection participants increases, so we considered all the reported privacy concerns valid. Instead, the tight-budget search focuses on a more realistic setting that practitioners may only have resources to address top privacy concerns. Since privacy opinions are often personal, some privacy concerns may only be reported by one individual. The tight-budget search considers a privacy concern valid if more than 2 out of 20 participants mentioned the same privacy concern for a data action.

Results: Figures 17 and 18 show the average proportion of privacy concerns found by each size of aggregate. These averages were calculated by a Monte Carlo technique, where we selected 10,000 random aggregates for each aggregate size and experiment.

Although individual inspectors only find few privacy concerns (tight-budget: 21.3% ± 5.7%; explorative: 13.3% ± 2.2%), collectively, aggregating their inspection results can reach good coverage. Aggregating 5 participants can find 66.3% of the privacy concerns in a tight budget

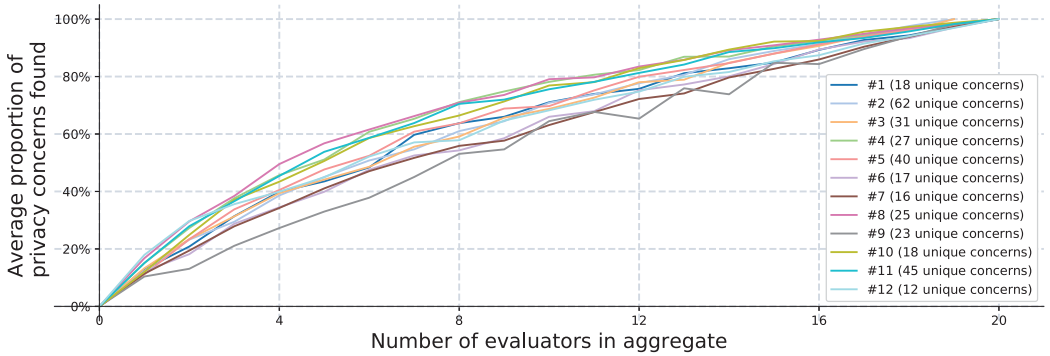


Fig. 17. Proportion of privacy concerns found by aggregates of size 1–20 in the explorative search for scenarios #1–#12. The explorative search analysis studies the discovery of privacy concerns saturation in an idealistic setting when the practitioners want to find any potential privacy concerns. Here, we considered all the reported privacy concerns valid. On average, 18 participants can find 95.5% (std = 1.4%) of privacy concerns.

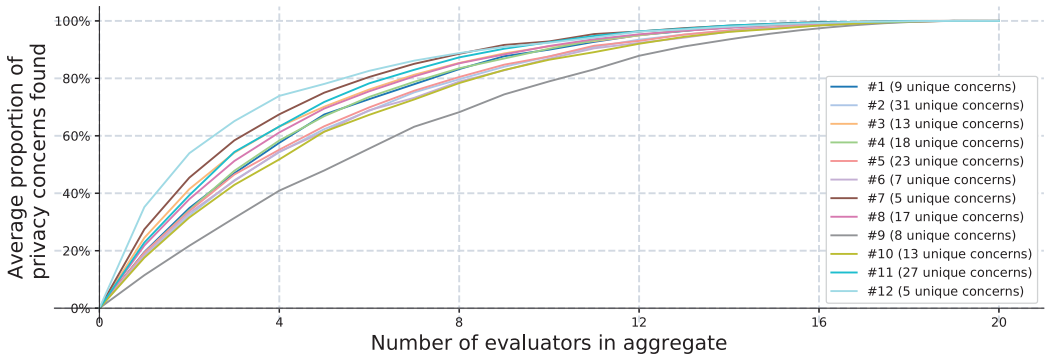


Fig. 18. Proportion of privacy concerns found by aggregates of size 1–20 in the tight-budget search for scenarios #1–#12. The tight-budget search focuses on a more realistic setting that practitioners may only have resources to address top privacy concerns. The tight-budget search considers a privacy concern valid if more than 2 out of 20 participants mentioned the same privacy concern for a data action. On average, 14 participants can find 97.1% (std = 1.2%) of privacy concerns.

search and 46.0% (std = 6.3%) of the privacy concerns in an explorative search. If we increase the number of participants to 14, the coverages will reach 97.1% (std = 1.2%) and 84.0% (std = 4.4%), respectively. We set the 95% coverage as the saturation cutoff threshold. The discovery of privacy concerns saturates at roughly 14 participants in the tight budget search, and 18 participants in the explorative search.

9.3.2 Cost and Latency. Here, we analyze the cost and latency of LPR to illustrate the benefits of crowd privacy inspection. A formal privacy review often requires a few weeks' turnaround time [13] and costs $10,000–$60,000 in human labor for companies [82]. Our experiments show that LPR can offer a report of privacy concerns at a small fraction (≈ 100X reduction) of the cost and wait-time for a formal review.

Results: Table 10 enumerates the cost breakdown of each data practice. On average, we spent around $114.45 for one privacy inspection, including $51.67 paid to inspectors (i.e., 4.52 hours of total crowd work), $36.95 paid to annotators (i.e., 0.76 hours of total crowd work), and $25.83

Table 10. The Cost Break Down of Tested Scenarios

	# of data actions	Inspection cost	Annotation cost	Hours of crowd work	$Cost_1$	$Cost_2$
1	7	($2.00 + $1.00) x 20	$0.20 x 3 x 48	3.86 + 0.61 (hours)	$114.80	$80.36
2	10	($2.00 + $1.00) x 20	$0.20 x 3 x 117	5.48 + 1.41 (hours)	$177.08	$123.96
3	6	($2.00 + $1.00) x 20	$0.20 x 3 x 60	5.55 + 0.76 (hours)	$123.80	$86.66
4	6	($2.00 + $1.00) x 20	$0.20 x 3 x 63	6.08 + 0.66 (hours)	$126.05	$88.24
5	5	($1.00 + $1.00) x 20	$0.20 x 3 x 74	3.87 + 0.91 (hours)	$106.90	$74.83
6	6	($2.00 + $1.00) x 20	$0.20 x 3 x 34	4.81 + 0.41 (hours)	$104.30	$73.01
7	4	($1.00 + $1.00) x 20	$0.20 x 3 x 40	2.73 + 0.51 (hours)	$81.40	$56.98
8	4	($1.00 + $1.00) x 20	$0.20 x 3 x 66	3.56 + 0.83 (hours)	$100.90	$70.63
9	4	($1.00 + $1.00) x 20	$0.20 x 3 x 24	3.70 + 0.32 (hours)	$69.40	$48.58
10	3	($1.00 + $1.00) x 20	$0.20 x 3 x 39	3.16 + 0.49 (hours)	$80.65	$56.46
11	6	($2.00 + $1.00) x 20	$0.20 x 3 x 130	5.73 + 1.67 (hours)	$176.30	$123.41
12	8	($2.00 + $1.00) x 20	$0.20 x 3 x 44	5.74 + 0.57 (hours)	$111.80	$78.26
Avg	5.75	$51.67	$36.95	4.52 + 0.76 = 5.28 (hours)	$114.45	$80.11

Since the actual cost also depends on the task reward, we also reported the total hours of crowd work.
Inspection cost = (Worker payment + Guaranteed bonus payment) x 20 Workers.
Annotation cost = (Reward per assignment + Number of assignments per annotation) x Number of annotations.
Transaction fee (not included) = TurkPrime Fee + Inspection Amazon turk fee + Annotation Amazon turk fee.
Hours of crowd work = Time to answer surveys + Time to annotate responses.
$Cost_1$ (The explorative search)= Inspection cost + Annotation cost + Transaction fee.
$Cost_2$ (The tight-budget search) = $Cost_1 \times 0.7$.

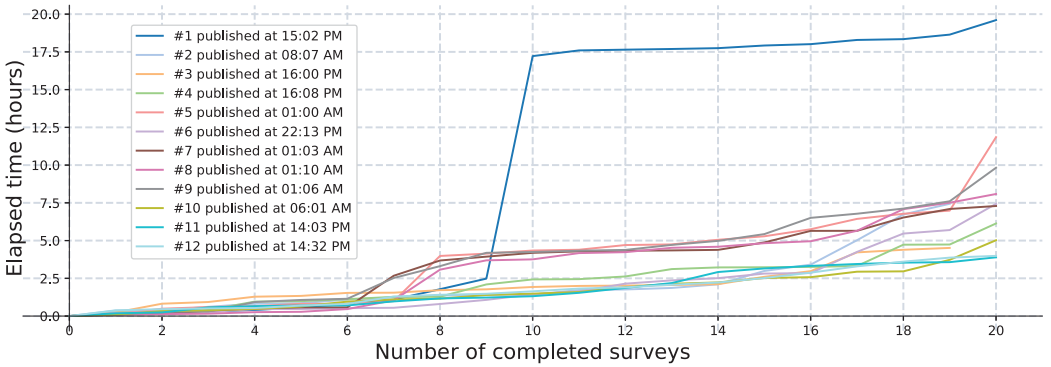


Fig. 19. Inspection latency for scenarios #1–#12. Completing 20 surveys takes around 5 hours since the process is not entirely parallel. The large jump in #1 is because we tested the "Microbatch" feature in TurkPrime, which runs the HITs in sequence. The main latency comes from the locking mechanism in AMT. When a participant claims the task but does not finish, another interested participant needs to wait until the crowdsourcing platform releases the lock.

paid to the platforms. If we only involve 14 participants in the study, the average cost would be around $80 (i.e., 3.70 hours of total crowd work). In general, the cost to inspect a data practice increases as it involves more data actions. Scenario #2 is the most expensive data inspection, which also contains the most data actions (N = 10). A negative data practice is more costly than a positive data practice since it receives more negative responses, resulting in more annotation tasks.

Figure 19 enumerates the survey completion times after releasing the complete set of tasks. The experiment for story #1 consumes more time since we tested the "Microbatch" feature in

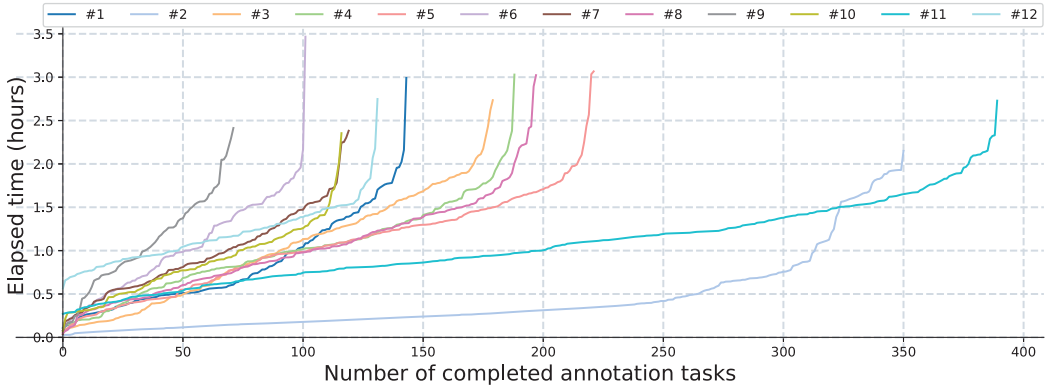


Fig. 20. Annotation latency. Annotation tasks are more parallel than the inspection tasks for two reasons: (1) each task takes less time to complete; and (2) a participant is allowed to complete multiple tasks. However, participants are less motivated to complete the last few leftover tasks (usually one or two), which results in the extended tail distribution in elapsed time. Fortunately, LPR can still identify the annotation by merging the available annotations together without these few leftover ones. So practitioners need to wait about 2 hours to get the free-text response annotations.

TurkPrime, which runs the HITs in sequence. The rest experiments all enabled the "Hyperbatch" feature, which ensures the fastest possible data collection by making the study available to all workers who may want to take it at the same time.

Figures 19 and 20 show the latency of inspection and annotation tasks. On average, it takes 3.30 (std = 1.16) hours to collect 14 inspection survey responses and 2 hours to annotate the responses, resulting in a total of latency of 5.5 hours.

9.3.3 Result Quality. The crowd inspection outputs both the **range** and **magnitude** of different privacy concerns. We evaluate these two aspects by comparing the inspection results with the ground truth list.

Method (coverage). To evaluate the coverage quality, we measure the number of valid privacy concerns (Valid), the number of missed privacy concerns (Miss), and the number of FA for each story. We considered a privacy concern valid if the privacy experts did not mark it as a FA. We considered a privacy concern missed if the privacy concern is only discovered by the other inspection method but it was not labeled as a FA by the judges.

Results (coverage). Table 11 enumerates the raw numbers of Valid, Miss, and FA for both LPR and the baseline approach. In total, LPR finds 315 valid privacy concerns, misses 38 privacy concerns, and reports 19 FA privacy concerns; while the baseline approach finds 138 valid privacy concerns, misses 215 privacy concerns, and reports 12 FAs. LPR finds 123 privacy concerns that are also identified by data practitioners (89.1%), as well as 192 privacy concerns that practitioners are not aware of (139.1%). The estimated FA rate for LPR and practitioners are 6.0% and 8.7%, respectively. LPR outperforms the baseline approach in all three metrics.

Method (magnitude). We define the magnitude of a privacy concern as the sum of associated comfortableness scores. For example, if two inspectors report a same privacy concern with "extremely uncomfortable" ratings (5), the magnitude of that privacy concern would be 10. For each data action, both approaches output a sorted list of privacy concerns based on the magnitude. We measure the **Normalized Discounted Cumulative Gain** (**nDCG**) [131] of each list to quantify the ranking quality.

Table 11. The Privacy Concern Coverage across 12 Stories

Participants	Story #1			Story #2			Story #3			Story #4		
	Valid	Miss	FA	Valid	Miss	FA	Valid	Miss	FA	Valid	Miss	FA
Crowd workers	**17**	**5**	**1**	**60**	**5**	2	**30**	**0**	1	**26**	**4**	1
Practitioners	13	9	3	20	45	**0**	6	24	1	12	18	**0**

Participants	Story #5			Story #6			Story #7			Story #8		
	Valid	Miss	FA	Valid	Miss	FA	Valid	Miss	FA	Valid	Miss	FA
Crowd workers	**37**	**4**	3	**15**	**0**	2	**15**	**2**	**1**	**23**	**1**	2
Practitioners	16	25	**2**	2	13	**1**	6	11	2	12	12	2

Participants	Story #9			Story #10			Story #11			Story #12		
	Valid	Miss	FA	Valid	Miss	FA	Valid	Miss	FA	Valid	Miss	FA
Crowd workers	**18**	**2**	5	**17**	**4**	1	**45**	**9**	0	**12**	**2**	0
Practitioners	7	13	**1**	7	14	**0**	28	26	0	9	5	0

A privacy concern is an FA if the privacy researchers think it non-relevant. We consider a privacy concern valid if researchers did not mark it as an FA, and missed if the concern is only discovered by the other inspection method. LPR outperforms the baseline approach in all three metrics. We bold the numbers to highlight the winning method.

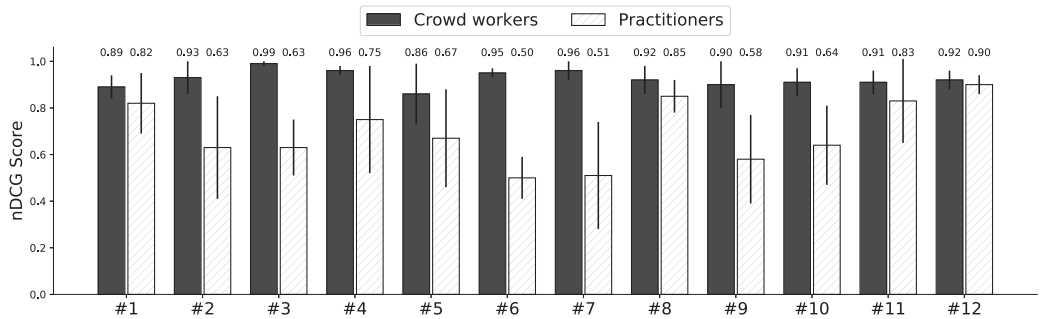


Fig. 21. Normalized discounted cumulative gain (nDCG). If the predicted list is perfectly aligned with the ground truth list, the nDCG equals 1. A 0.8 nDCG indicates 80% of the best ranking. The privacy concern output from crowd workers outperforms the output from practitioners in all scenarios.

$$nDCG_p = \frac{DCG_p}{DCG_m} \quad where \quad DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}, \tag{1}$$

where DCG_p is the **Discounted Cumulative Gain** (**DCG**) for a particular rank p, DCG_m is the DCG for an ideal order (i.e., the ground truth order), i is the order of a specific privacy concern in the output list, and rel_i is ground truth priority value for the privacy concern. Indeed, nDCG is the ratio of the DCG of a recommended order to the DCG of the ideal order. If the predicted list is perfectly aligned with the ground truth list, the nDCG equals 1. An 0.8 nDCG indicates 80% of the best ranking.

Results (magnitude). We compute the nDCG scores for each data action and aggregate the scores of actions in the same story. Figure 21 enumerates the nDCG scores for all 12 stories. The privacy concern output from crowd workers achieves an average nDCG of 0.925, which outperforms the output from practitioners (0.693).

To gain a better understanding of the nDCG values, we present the magnitudes of the concerns associated with a data action in Table 12. As we can see, the results from crowd workers are

Table 12. The Magnitude of Privacy Concerns for the "#7 Score Manipulation" Action in Appendix Figure 29

Source	The magnitude of privacy concerns	DCG	nDCG
Ground truth	Violation of expectations (28), Deceptive data practice (25), Lack of informed consent (23), Lack of alternative choice (23), Lack of trust for algorithms (22), Lack of control of personal data (22), and Insufficient data security (12)	101.27	1
Crowd workers	Deceptive data practice (52), Lack of trust for algorithms (28), Lack of informed consent (10), Violation of expectations (5), Lack of control of personal data (5), and Insufficient data security (5)	95.67	0.945
Practitioners	Deceptive data practice (19), Lack of alternative choice (5), and Data commodification (4)	77.68	0.767

The results from crowd workers are closely aligned with the assessment by privacy researchers with some minor mismatches, resulting in an nDCG of 0.945. Practitioners correctly found the top privacy concerns, missed four privacy concerns, and reported an FA, producing an nDCG of 0.767.

closely aligned with the assessment by privacy researchers, with two minor mismatches: (1) crowd workers miss a low priority concern—lack of alternative choice; and (2) crowd workers assign a wrong priority to "violation of expectations," resulting in an nDCG of 0.945. On the other side, practitioners find the top privacy concerns correctly, miss four privacy concerns, and report an FA, producing an nDCG of 0.767.

10 DISCUSSION

10.1 When and How to Use LPR?

Privacy feedback from actual users is often only collected after the software updates have been deployed (e.g., [76]). LPR makes it easier to collect user feedback about a data practice more frequently throughout the development lifecycle, similar to heuristic evaluation for usability feedback. We envision that the primary use cases for LPR are to (1) detect possible backlash over data collection/sharing/processing/usage when a formal privacy review is not available and (2) explore alternative benign data practice variants through storyboard iterations.

One potential application of LPR is the integration of agile software engineering [10]. For example, a product manager may turn an ongoing "epic" or new "user story"[17] into a privacy storyboard, bring the results to an all-hands meeting discussion, and discuss privacy implications with the team. Past research [110, 135] shows that minor changes to a data practice design can significantly reduce users' disapproval and concern. However, exploring alternative benign variants through formal privacy reviews is challenging. The low-cost and lightweight nature of LPR can allow practitioners to iteratively improve their data practices based on fast user feedback, which can help a product team in avoiding implementing problematic software features/data practices.

Further, LPR can be used with great flexibility. First, practitioners can build knowledge of the data practice incrementally. A practitioner might not have all the required knowledge for a complex data practice. In that case, she can build storyboards for the familiar parts as a starting point, and incrementally add and modify data actions as needed. The design of the tree-typology and the decoupling of risk–benefit analysis support these usages. Second, the number of crowd workers involved is flexible. Our result shows that 14 participants found 97% of privacy concerns. However,

[17]"Epic" and "user story" are terms of agile software development, which refer to task narratives for sprint planning [8].

practitioners may have fewer participants for early prototypes (e.g., refining wordings) and more participants for later stage prototypes, along with full privacy reviews.

10.2 Crowd Worker Engagement

Informally, crowd workers felt positive about the participation experience. At the end of all the crowd surveys, we asked participants to provide optional feedback regarding the task experience. Most participants commented favorably. One said: "This was one of the best surveys I've done in a long time. Very thought-provoking." Another said: "Thank you for the interesting survey. I enjoyed reading and reflecting on my feelings for each scenario." In addition, the average length of free-text responses exceeded (213.3 characters) the required length (100 characters), and a few participants enthusiastically wrote a multi-paragraph composition to elaborate on their concerns.

10.3 The Actual Cost and Latency of Conducting an LPR

Note that the actual cost and latency of reviewing a specific data practice depends on multiple factors, such as the task reward, the number of data actions, and the crowd worker market's engagement. Instead of predicting the exact cost for future LPR reviews, our goal in this article is to demonstrate that the crowd inspection can be an order of magnitude cheaper and faster than a traditional privacy review, which often costs several thousand US dollars [83, 124] and has a turnaround time of a few weeks [59, 97].

In practice, practitioners can further reduce latency by releasing extra tasks. The main latency in our experiments comes from the locking mechanisms in AMT (see Figure 20): when a participant claims the task but does not finish, another interested participant needs to wait for the crowdsourcing platform to release the lock. By releasing extra tasks, practitioners can collect enough responses without waiting for the last few locked tasks, although it will cost extra money.

11 LIMITATIONS: WHEN DOES LPR NOT WORK?

11.1 Requirements for Practitioners

LPR imposes two requirements for practitioners. First, practitioners need to communicate the data practice in an LPR story honestly. Since participants in LPR understand the data practice using the privacy storyboard rather than through real-world interactions, it is possible that different wordings and framings in the descriptions may impact their perception [15]. Techniques avoiding self-deception in business ethics [99] might alleviate this effect.

Second, practitioners need to have a comprehensive understanding of the data practice design and communicate it accurately to the participants. For example, when reviewing the Uber battery-based privacy surging data practice, a machine learning expert may challenge if the black-box machine learning model stealthily optimized the surge price based on users' smartphone battery data. However, crowd workers lack the expertise to foresee these types of issues. It would be the practitioners' responsibility to run extra sanity tests to understand the implications. Since LPR allows users to build and modify the storyboard incrementally, practitioners can potentially build such an understanding through multiple iterations.

11.2 Crowd Worker Representativeness

Our experiments in this article approximate the everyday user population by querying AMT crowd workers. On-demand crowd workers can provide privacy-relevant feedback in a faster and cheaper way than dedicated privacy specialists (e.g., experienced privacy engineers and UX researchers). However, such an approximation may not resemble the intended user population for a given product or service. One potential solution to address this issue is to consider recruiting participants

from multiple channels (e.g., marketing survey for real users and internal dog food users) or direct LPR to the targeted user group of the specific data practice in mind.

In our evaluation, all the crowd workers can only participate our study once to avoid the learning effect. However, in practice, if crowd workers get more experienced in LPR surveys over time, they might better calibrate across data practices and spot common blind spots. Future work may look into methods to improve the productivity of crowd workers on privacy inspection.

12 FUTURE WORK

12.1 Understanding the Iteration Process of Data Practice Design

Our current study focuses on the feasibility and benefits of crowd inspection, which has not explored the iterative process of data practice designs. Future research should try to run multi-session studies with practitioners to observe how practitioners understand crowd workers' responses, derive new hypotheses for the sources of privacy concerns, and develop new data practice variations to validate their assumptions.

12.2 Privacy Training for Practitioners through LPR

Privacy training for practitioners has been challenging [28]. The current formal privacy review paradigm may perpetuate an undesirable trend: data practitioners tend to leave privacy considerations to dedicated specialists and do not think about them actively. In contrast, the design of LPR requires practitioners to be in the loop of the privacy decision-making process, allowing them to access the actual feedback from users and test their hypotheses at a low cost. Future research should run field studies to test whether practitioners' can better expect users' privacy concerns after using LPR, and whether this acquired privacy expertise are transferrable to unseen data practices.

12.3 Guidelines and Best Practices for Authoring Storyboards

To ensure quality storyboard, we asked three participants with some background in privacy to create the storyboards (see Section 8.1). We found that participants can create storyboards in 20–30 minutes. However, similar to UX storyboarding [121], novice users may encounter various challenges in the creation of storyboards (e.g., wording and flow organization). Future work should look into characterizing what kinds of privacy storyboards are most effective and what kinds of guidelines and tools (beyond our worksheet) can best help practitioners create those storyboards.

12.4 Summarizing Positive Opinions

The current design of LPR focuses on the negative opinions and can only summarize them into the magnitude and range of privacy concerns. For positive opinions, practitioners can only read the raw text indexed by comfortableness scores and users (e.g., Figures 10 and 11). Future work should look into building a taxonomy for the positive opinions and building interfaces to help users contrast the positive opinions with the negative ones on the same topics.

12.5 Creating New Privacy Indexes to Educate Users and Inform Companies

It is possible to use LPR to collect privacy opinions from a large number of participants on exemplar data practices, making it possible to build a new type of privacy index that can quantify the ranges and magnitudes of privacy concerns for different demographics. We can use such an index in two ways. First, a consumer may compare their responses to the index to better calibrate their privacy sensitivity [31, 126]. For example, LPR can show a participant that 30% of other participants share the same privacy concerns of you, and 20% reported different privacy concerns. Second, the index

can potentially depict a privacy norm for a specific data practice design in the form of aggregated privacy concerns (i.e., range and magnitude) quantitatively.

13 CONCLUSION

We introduce LPR, a fast, cheap, and accessible system to help practitioners evaluate users' privacy concerns of data practices in consumer-facing businesses. LPR takes a proposed data practice, quickly breaks it down into smaller parts, generates a set of questionnaire surveys, solicits users' opinions through the proxy of crowd workers, and summarize those opinions in a compact form for practitioners to use. Our experiments show that LPR finds 89% of privacy concerns identified by data practitioners as well as 139% more privacy concerns that practitioners are not aware of, at a 6% estimated FA rate.

LPR contributes three unique benefits to existing privacy solutions. First, LPR is more accessible than existing technologies, which can help small teams and individuals who do not have resources to hire dedicated privacy specialists. Second, LPR can help professional teams collect direct feedback from end users, ameliorating potential bias and blind spots. Third and more broadly, LPR advocates users' voices in the increasingly important privacy debate.

APPENDIX

A COMPLETE LPR PRIVACY CONCERN TAXONOMY

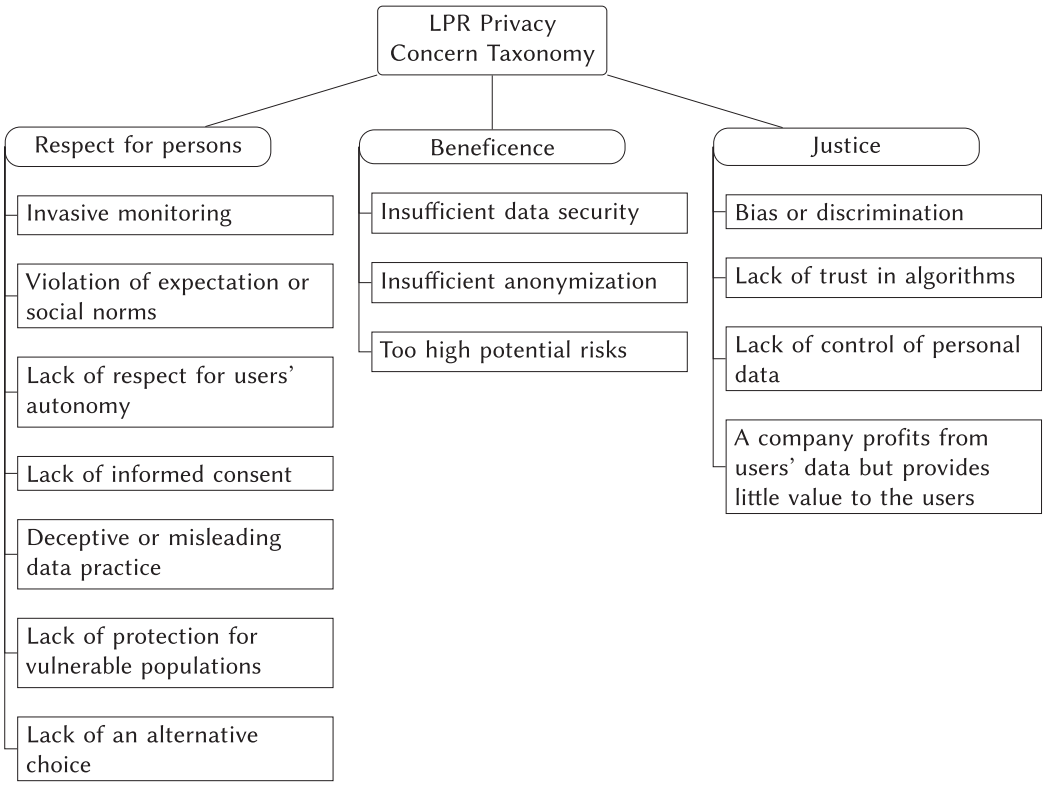


Fig. 22. The overview of LPR privacy concern taxonomy. We used a ground theory approach to build the list of common privacy concerns, which is the basic vocabulary to characterize users' free-text privacy responses. To help practitioners comprehend the list quickly, we further organized these privacy concern types into three high-level categories: respect for persons, beneficence, and justice.

Table 13. The **Privacy Concerns** (**PC**) under *Respect for Persons*

PC category	Example subcategories	Example quotes
Invasive monitoring	Too much data	I feel uncomfortable because I am being watched so closely.
	Too sensitive data	I don't like the idea of my fertility data being used in any way.
Violation of expectations/social norms	Unexpected data sharing	No one should be sharing this with other companies unless there was illegal activity going on.
	Unexpected data collection	I don't think the company should have any information like that.
	Unexpected data appropriation	I don't think that how someone acts in a game should impact their ability to get a job.
Lack of respect for autonomy	Decisional interference	Marketers are using this information to get me to click on their ads, and it may not be the product that draws my attention but rather the color.
	Users prefer self-control than data-driven automation	Just let me select the item and pay for it. I don't need these false conveniences.
Lack of informed consent	Lack of transparency	The app is now operating in hidden ways that are not explained to the user.
	Lack of consent	I don't like knowing they are using psychological tactics to gain profit from me based on info they acquired without my consent.
	Violation of existing consent	I don't think that I agree to share my data.
Deceptive or misleading data practice		I feel uncomfortable because the company is misleading me, and that goes against the tacit understanding I had when agreeing to use the app.
Lack of protection for vulnerable populations		I think the company should not send pregnancy product ads to high school girls.
		I feel targeting ads at sad/anxious teenagers is unethical.
Lack of an alternative choice.	No opt-out option	I feel uncomfortable because I cannot choose not to participate.

Respect for persons: the company should respect users' dignity, autonomy, and provides special protection to vulnerable populations. The example quotes are from crowd workers' survey responses.

Table 14. The PC under *Beneficence*

PC category	Example subcategories	Example quotes
Insufficient data security	Potential data breach	I feel uncomfortable because a data breach may occur.
Insufficient anonymization	Insufficient anonymization	Because my company is small, so it wouldn't be very anonymous even without identifiable information.
Too high potential risks	Risk of financial loss	I don't trust that the company won't just use the data to further cut their own costs whenever possible and further screw over their own employees with less benefits, etc.
	Risk of opportunity loss	I don't want my private sensitive health information revealed to anyone really, especially my employers, when it may impact employee promotions or other decisions.
	Risk of reputation loss	

Beneficence: the benefits of a data practice for users should justify the potential risks/costs. The example quotes are from crowd workers' survey responses.

Table 15. The PC under *Justice*

PC category	Example subcategories	Example quotes
Bias or discrimination		The idea of fluctuating pricing seem somewhat unethical and possibly illegal. Everyone should be charged the same.
Lack of trust for algorithms	Concerns on the imperfect implementation	I would not trust their algorithms.
	Concerns on algorithmic automation	I don't like the idea of a store getting ideas about me based solely on what I buy. I think I'm more dimensional than my purchases alone may suggest.
Lack of control of personal data		I have no way to control the data after sharing the data.
A company profits from users' data but provides little value to the users (i.e., data commodification)		The company will use this data to profit off my shopping and buying habits. I do not see any benefit of this save for additional targeted marketing.
		I feel uncomfortable because I don't want my clicks tracked for the purpose of making a company more money off of me through ads.

Justice: each user deserves equal and fair treatment in the given data practice. The example quotes are from crowd workers' survey responses.

Table 16. There Can Be Three Types of Data Stakeholders: Data Subjects, Data Observers, and Data Beneficiaries/Victims

Primitives	Definitions	Examples
Data subject	the people who contribute their data	female employees
Data observer	the entities that have access to people's data	the App developer and the employers
Data beneficiaries/victims	the people impacted by a data practice	the employees, the app developer and the employers

From the news report above, we identify the following stakeholders.

B DATA ACTION ANALYSIS WORK SHEET

B.1 Overview

This worksheet aims to help a data practitioner construct a privacy storyboard based on a data practice he/she has in mind. A privacy storyboard contains a set of data actions organized in a tree topology. We provide definitions, instructions, and examples below.

B.2 An Example Data Practice

We illustrate the procedure of privacy storyboarding through an example data practice, derived from "Is your pregnancy app sharing your intimate data with your boss?" published in The Washington Post in April 2019.[18]

We offer a summary of the news report below:

The period- and pregnancy-tracking app Ovia helps users track their pregnancy journeys. Employers who pay the apps' developer can offer their workers a special version of the apps that relays their health data—in a "de-identified," aggregated form—to an internal employer website accessible by human resources personnel. The companies offer it alongside other health benefits and incentivize workers to input as much about their bodies as they can, saying the data can help the companies minimize health-care spending, discover medical problems, and better plan for the months ahead. However, experts worry that the employers could identify women based on information relayed in confidence, particularly in workplaces where few women are pregnant at any given time.

B.3 Instructions

Step 1(a). Enumerate the potential outcomes that will be generated by the data practice. The news report describes two possible outcomes (data applications) as follows:

(1) Employers can use the information to determine when and how many female employees will take maternity leave, which can then be used to plan for openings and health insurance.
(2) Although the pregnancy related data are anonymized, the employer may still be able to identify the specific female employee through other information.

In practice, it might be challenging to enumerate every possible outcome in one pass; the focus here should be on the most possible and common cases (both positive and negative). You may revisit this step later and add new outcomes incrementally.

Step 1(b). Identify the data stakeholders for each data application. A data stakeholder is an individual or group that could affect or be affected by the data practice. In practice, you may identify more outcomes after enumerating the stakeholders. You may revisit Step 1(a) to iterate the storyboard.

[18]See https://www.washingtonpost.com/technology/2019/04/10/tracking-your-pregnancy-an-app-may-be-more-public-than-you-think/?arc404=true\for\the\news\report.

Table 17. A Data Action is the Smallest Unit in an LPR Story, Which Describes a Specific Operation that Consumer Businesses Interact with Users' Data

Primitives	Definitions
Data collection	A data observer collects/stores data from data subjects.
Data processing	A data observer processes users' data to derive new data.
Data sharing	A data observer shares users' data or derived data with another data stakeholder (i.e., observer, subject, and beneficiary/victim).
Data usage	A data observer uses the data in a certain way that impacts a data beneficiary/victim.

We used four types of data action primitives: (1) Data collection; (2) Data sharing; (3) Data processing; and (4) Data usage.

Step 2(a). Identify the data actions in a data practice and organize them in a tree typology. We identify the following data actions from the news story above:

(1) Data collection: A pregnancy app collects data from their uses.
(2) Data processing: The app processes users' behavior data and generates some statistics.
(3) Data sharing: The app shares the data with many employers.
(4) Data usage: Employers use the data to determine healthcare spending.
(5) Data usage: Employers use the data to identify female employees who are pregnant.

Step 2(b). Organize the flow of the data actions. We organize the two data applications as the following two sequences with shared data actions.

Application 1: (1) -> (2) -> (3) -> (4)
Application 2: (1) -> (2) -> (3) -> (5)

Step 3. Describe each data action in succinct text. The last step is to generate a text description for each data action, explaining the details of a data action to the non-tech-savvy audience. An ideal description should contain all the necessary privacy-salient information (e.g., how the data are being collected, who it is being shared with, what information is being derived, and how the data are being used) while being succinct.

You may think each data action through the following questions and use a lightweight description structure: [Context] [Data action][Further description - optional].

Questions for data collection.

— **Context:** Who is the data observer? What is the data subject? What data are being collected/stored? Why is the company collecting/storing the data?
— **Data action:** How is the data being collected?
— **Further description:** Is the data collection anonymous/pseudo-anonymous/non-anonymous? Do the users receive an explicit notification? How will the company keep the data? When will they delete the data? Do users receive compensation?

Data action (1): A pregnancy app collects data from their users.
Detailed description: [Context] A pregnancy-tracking app provides health-care aid for women to understand their bodies better. Users can log in to record their bodily function, e.g., sex drive, medications, and mood. The app has a "fertility algorithms," which analyze the users' menstrual data and suggest good times to try to conceive. [Data action] Your company establishes a partnership with the app developer. [Further description] Your company pays each employee $1 a day in gift card if she regularly uses the app.

Questions for data processing.

- **Context:** Who is the observer? What are the raw data? What are the derived data?
- **Data action:** How are the data being processed?
- **Further description:** Is the processing anonymized? Who did the processing? Algorithms or humans? How is the processing method developed?

Data action (2): The app processes users' health data and generate related statistics.
Detailed description: The pregnancy-tracking app aggregates the data of employees in the same company and removes personally identifiable information (e.g., name, e-mail, and age).

Questions for data sharing.

- **Context:** Who is the sender? Who is the recipient? Why does the sender share the data?
- **Data action:** How are the data being shared?
- **Further description:** Is the data sharing for profit? Is the process secure?

Data action (3): The app shares the data with many employers.
Detailed description: The employer pays the app developer to get their employees' aggregated data and related statistics, e.g., how many workers using the app had faced high-risk pregnancies or had given birth prematurely; how soon the new moms planned to return to work.

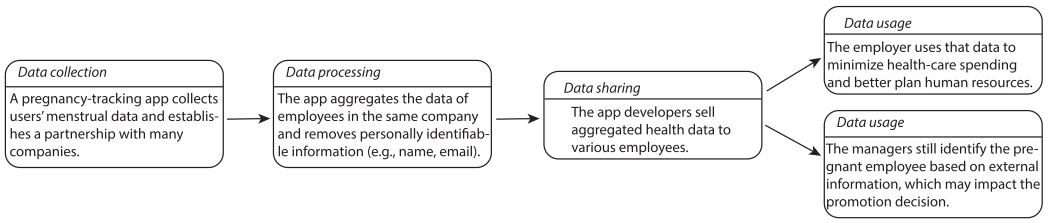
Questions for data usage.

- **Context:** Who is the observer? Who is the data beneficiary/victim?
- **Data action:** What are the potential risks (i.e., the probability and impact) for impacted users?
- **Further description:** Are users aware of such data usage? Does the company have users' informed consent? Does the company give users compensation?

Data action (4): Employers use the data to determine healthcare spending.
Detailed description: The employer plans to use the data to minimize health-care spending, discover medical problems, and better plan human resources for the months ahead.

Data action (5): Employers use the data to identify female employees who are pregnant.
Detailed description: The employer can still identify the employee based on information relayed in confidence, particularly in workplaces where few women are pregnant at any given time. These intimate information may impact that employee's promotion.

Here, we generate the final version of privacy storyboard for this data practice.



C 12 REAL-WORLD PRIVACY STORIES AND THEIR PRIVACY STORYBOARDS

Figures 23–34 illustrate the privacy storyboards of 12 real-world data practices. We only offer a brief text summary for each data action in these illustrations. The complete textual description is available on the accompanying website.

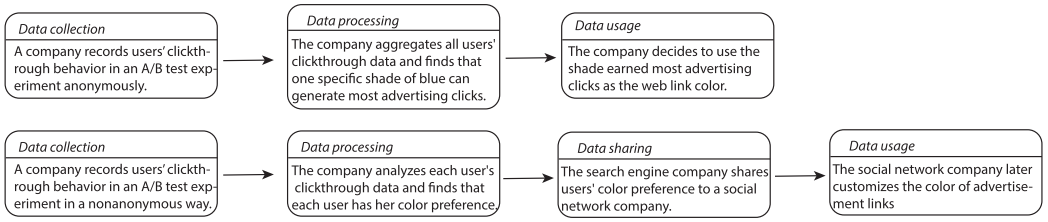


Fig. 23. A privacy storyboard of "#1 Search engine clickthrough data." A company records users' clickthrough behavior in an A/B test experiment anonymously and uses the data for advertising and search personalization.

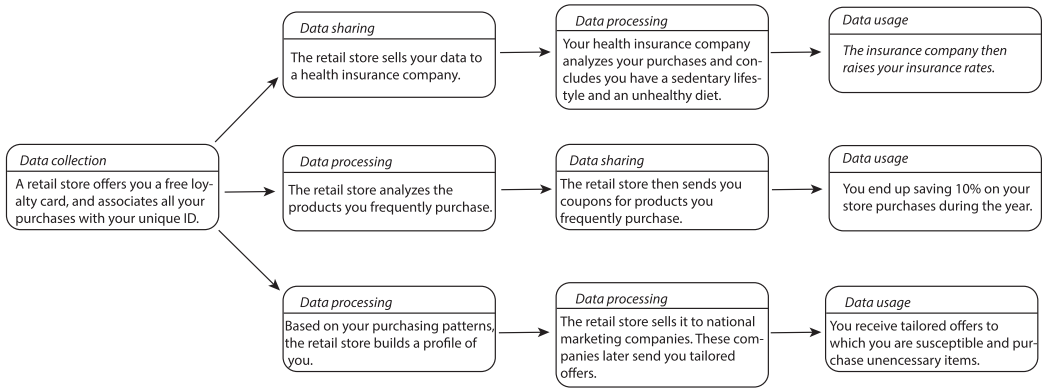


Fig. 24. A privacy storyboard of "#2 Loyalty card in a retail store." A retail store collects users' data through a loyalty card and uses the data for insurance and coupon personalization.

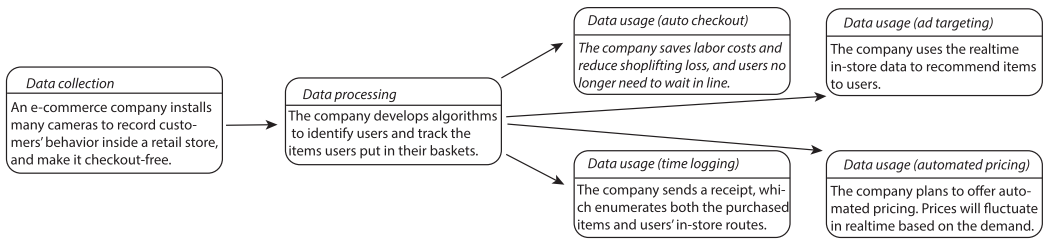


Fig. 25. A privacy storyboard of "#3 checkout free retail store." An e-commerce company opens a checkout-free retail store by installing various sensors inside a physical store.

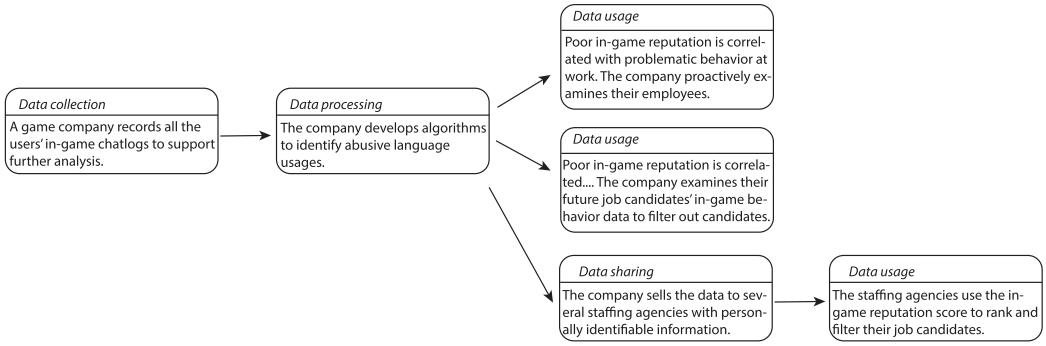


Fig. 26. A privacy storyboard of "#4 game chat log." An online game company uses chatlogs to identify potential problems in the workspace.

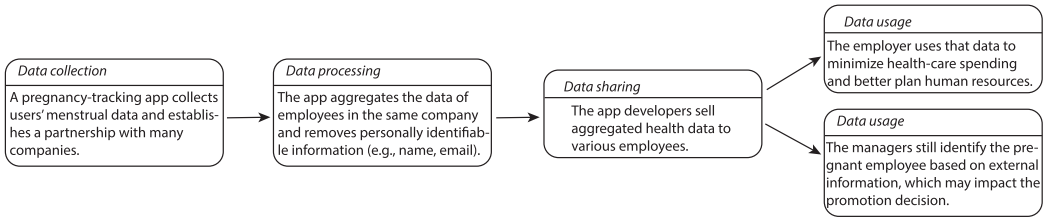


Fig. 27. A privacy storyboard of "#5 pregnancy intimate data." A pregnancy app shares users' intimate body data with their employers.

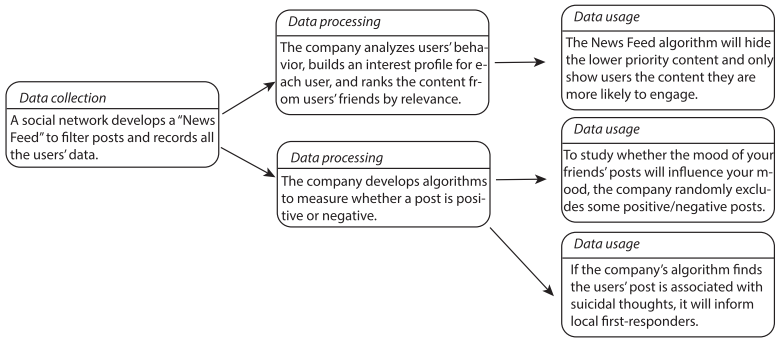


Fig. 28. A privacy storyboard of "#6 social network." A social networking service company analyzes users' posts through sentiment analysis and uses insights in different ways.

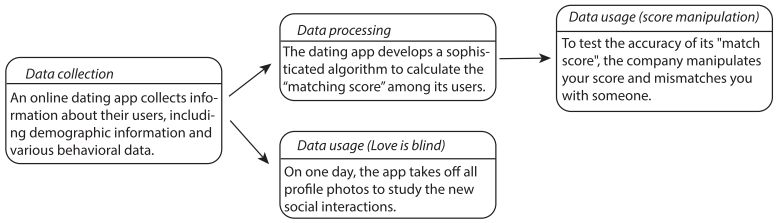


Fig. 29. A privacy storyboard of "#7 data science experiments in a dating app." An online dating app conducts several experiments to understand the nature of romantics.

Fig. 30. A privacy storyboard of "#8 Email contacts for social network bootstrapping." A technology company appropriates users' e-mail data to bootstrap a new social network service.

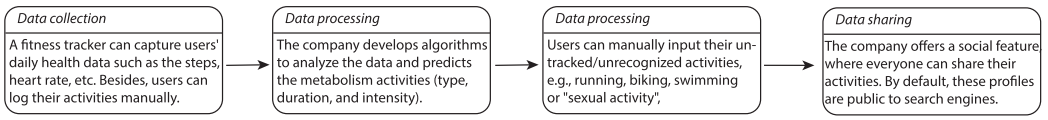


Fig. 31. A privacy storyboard of "#9 Fitness tracking." A wearable technology company collects users' intimate behavior data and makes them public by default.

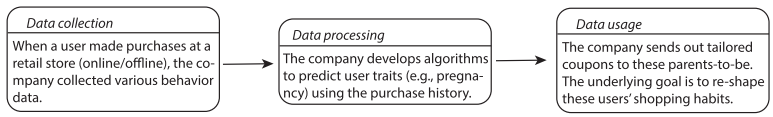


Fig. 32. A privacy storyboard of "#10 retail store pregnancy." A retail store predicts users' pregnancy status by analyzing their purchase history.

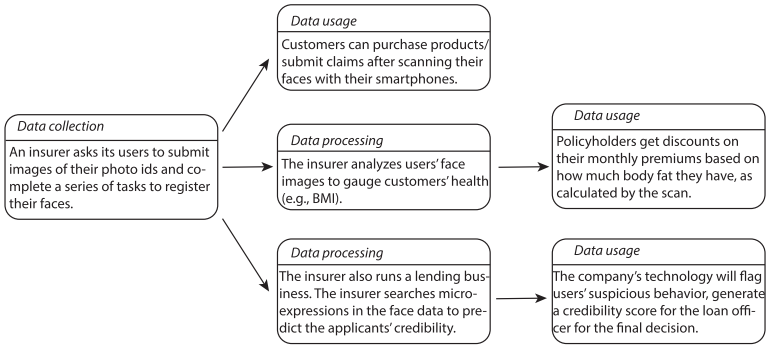


Fig. 33. A privacy storyboard of "#11 an insurer employs AI." An insurance company uses facial-recognition technology to identify untrustworthy and unprofitable customers.

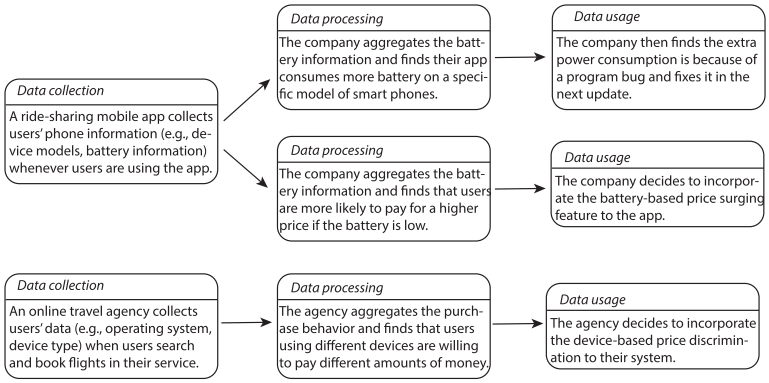


Fig. 34. A privacy storyboard of "#12 dynamic pricing." Technology companies collect users' behavior data to adjust the service price dynamically.

D THE DISTRIBUTION OF RESPONSES ACROSS DIFFERENT SCORES FOR INDIVIDUAL SCENARIOS

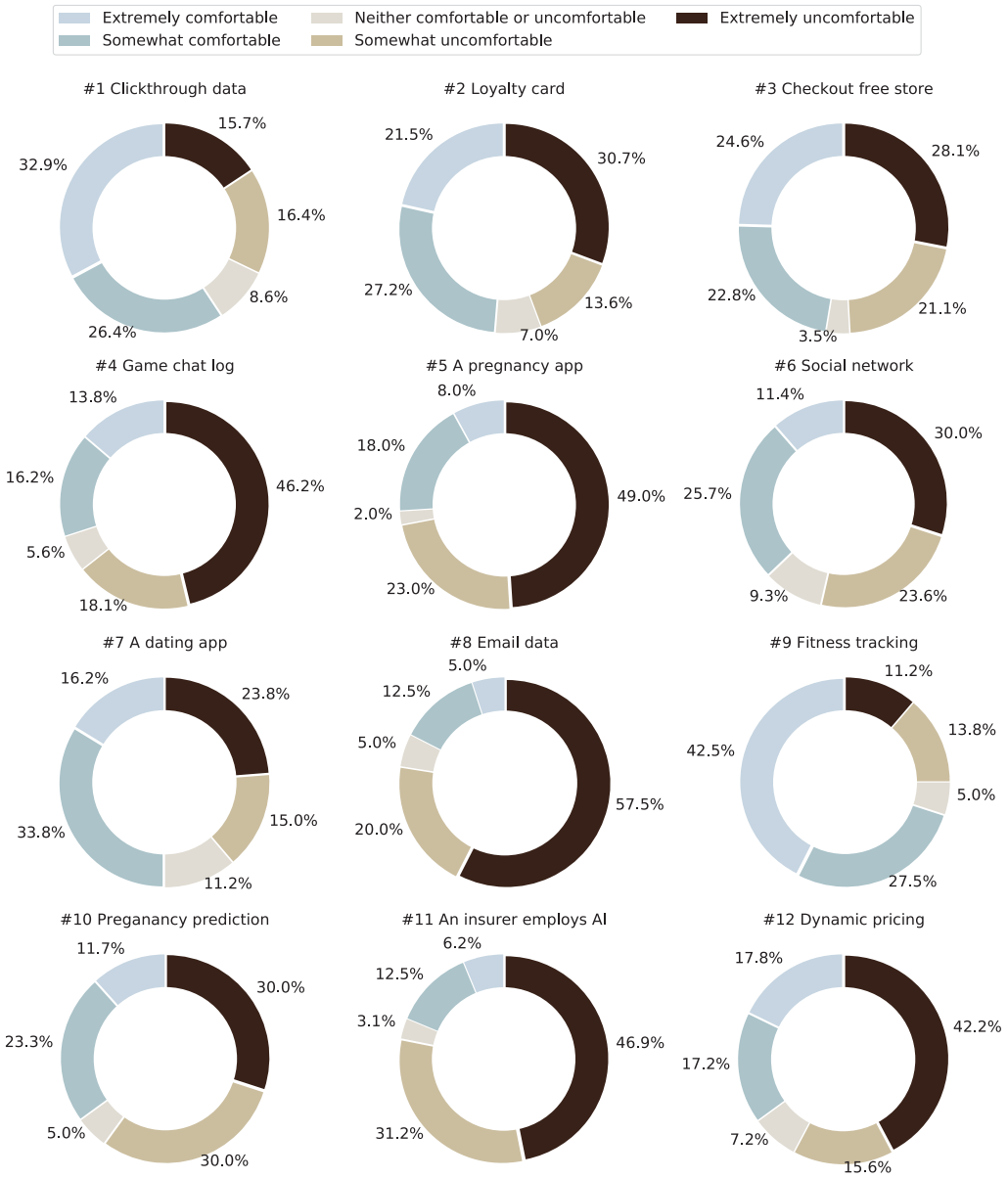


Fig. 35. The individual distribution of responses in each scenario across different scores. At a glance, the results offer many surprising results. For example, the seemingly less severed data practice, Google Buzz data appropriation [52, 88] (i.e., "#8 Email data"), received the most negative feedback. At the same time, the highly cited emotion contagion experiment [27, 65, 71, 104] (i.e., "#6 Social network") is less criticized by the crowd workers. Our focus of this article is to introduce the LPR framework; we will present the findings from our data in a separate document.

REFERENCES

[1] Roger Clarke. 2016. Privacy Introduction and Definitions. Retrieved July 17, 2018 from http://www.rogerclarke.com/DV/Intro.html.

[2] Fox Van Allen. 2014. When the Device You Use Determines the Price You Get. *Techlicious*. Retrieved from https://www.techlicious.com/blog/price-discrimination-by-operating-system-device/.

[3] Amazon. 2016. Does Amazon Have a Minimum Character Requirement for Reviews?—Selling on Amazon / General Selling Questions—Amazon Seller Forums. Retrieved February 17, 2020 from https://sellercentral.amazon.com/forums/t/does-amazon-have-a-minimum-character-requirement-for-reviews/130681.

[4] Cynthia Van Ort, Andrew Clearwater, and Chad Quayle. 2016. An Agile Approach to PIAs and Privacy by Design. Retrieved March 30, 2020 from https://iapp.org/resources/article/an-agile-approach-to-pias-and-privacy-by-design/.

[5] Annie I. Antón, Julia B. Earp, and Jessica D. Young. 2010. How internet users' privacy concerns have evolved since 2002. *IEEE Security & Privacy* 8, 1 (2010), 21–27.

[6] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2, Article 59 (July 2018), 23 pages. DOI: https://doi.org/10.1145/3214262

[7] Charles Arthur. 2012. Marissa Mayer's appointment: What does it mean for Yahoo? | Technology | The Guardian. Retrieved January 30, 2019 from https://www.theguardian.com/technology/2012/jul/16/marissa-mayer-appointment-mean-yahoo?newsfeed=true.

[8] Atlassian. 2020. Epics, Stories, Themes, and Initiatives | Atlassian. Retrieved April 29, 2020 from https://www.atlassian.com/agile/project-management/epics-stories-themes.

[9] Louise Barkhuus. 2012. The mismeasurement of privacy: Using contextual integrity to reconsider privacy in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 367–376. DOI: https://doi.org/10.1145/2207676.2207727

[10] Kent Beck, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas 2001. Manifesto for agile software development. https://agilemanifesto.org/.

[11] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*. ACM, 313–322.

[12] Jaspreet Bhatia and Travis D. Breaux. 2018. Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction* 25, 6, Article 34 (December 2018), 47 pages.

[13] Engin Bozdag. 2020. Privacy at Speed: Privacy by Design for Agile Development at Uber. USENIX Association, San Francisco, CA.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[15] Alex Braunstein, Laura Granka, and Jessica Staddon. 2011. Indirect content privacy surveys: Measuring privacy without asking about it. In *Proceedings of the 7th Symposium on Usable Privacy and Security*. ACM, 15.

[16] Sean Brooks, Sean Brooks, Michael Garcia, Naomi Lefkovitz, Suzanne Lightman, and Ellen Nadeau. 2017. *An Introduction to Privacy Engineering and Risk Management in Federal Systems*. US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.

[17] Ryan Calo. 2013. Digital market manipulation. *George Washington Law Review* 82, 4 (2013), 995.

[18] Nicholas Carlson. 2010. WARNING: Google Buzz Has a Huge Privacy Flaw—Business Insider. Retrieved December 18, 2018 from https://www.businessinsider.com/warning-google-buzz-has-a-huge-privacy-flaw-2010-2.

[19] Ann Cavoukian. 2011. *Privacy by Design in Law, Policy and Practice. A White Paper for Regulators, Decision-Makers and Policy-Makers*. Information and Privacy Commissioner, Ontario.

[20] Bill Chappell. 2013. Google: Don't Expect Privacy When Sending to Gmail | Technology | The Guardian. Retrieved April 11, 2020 from https://www.theguardian.com/technology/2013/aug/14/google-gmail-users-privacy-email-lawsuit.

[21] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1999–2008.

[22] Amit Chowdhry. 2016. Uber: Users Are More Likely to Pay Surge Pricing if Their Phone Battery is Low. *Forbes*. Retrieved from https://www.forbes.com/sites/amitchowdhry/2016/05/25/uber-low-battery.

[23] David Cohen, Mikael Lindvall, and Patricia Costa. 2004. An introduction to agile methods. *Advances in Computers* 62, 03 (2004), 1–66.

[24] Federal Trade Commission Staff. 2012. Protecting consumer privacy in an era of rapid change–A proposed framework for businesses and policymakers. *Journal of Privacy and Confidentiality* 3, 1 (2011).

[25] The Nielsen Company. 2011. Privacy Please! U.S. Smartphone App Users Concerned with Privacy When it Comes to Location. Retrieved December 20, 2018 from https://www.nielsen.com/us/en/insights/news/2011/privacy-please-u-s-smartphone-app-users-concerned-with-privacy-when-it-comes-to-location.html.

[26] Sunny Consolvo, Ian E. Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. 2005. Location disclosure to social relations: Why, when, & what people want to share. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 81–90.

[27] Josh Constine. 2017. Facebook Rolls Out AI to Detect Suicidal Posts before They're Reported | TechCrunch. Retrieved February 04, 2020 from https://techcrunch.com/2017/11/27/facebook-ai-suicide-prevention/?guccounter=1.

[28] Lorrie Faith Cranor and Norman Sadeh. 2013. Privacy engineering emerges as a hot new career. *IEEE Potentials* 32, 6 (2013), 7–9.

[29] Mary J. Culnan and Pamela K. Armstrong. 1999. Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science* 10, 1 (1999), 104–115.

[30] Adele Da Veiga. 2018. An information privacy culture instrument to measure consumer privacy expectations and confidence. *Information & Computer Security* 26, 3, (2018), 338–364.

[31] Sauvik Das. 2016. Social cybersecurity: Understanding and leveraging social influence to increase security sensitivity. *it-Information Technology* 58, 5 (2016), 237–245.

[32] Scott Davidoff, Min Kyung Lee, Anind K. Dey, and John Zimmerman. 2007. Rapidly exploring application design through speed dating. In *Proceedings of the International Conference on Ubiquitous Computing*. Springer, 429–446.

[33] George Demiris, Brian K. Hensel, Marjorie Skubic, and Marilyn Rantz. 2008. Senior residents' perceived need of and preferences for "smart home" sensor technologies. *International Journal of Technology Assessment in Health Care* 24, 1 (2008), 120–124.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[35] Department of Health, Education, and Welfare; National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 2014. The Belmont report. Ethical principles and guidelines for the protection of human subjects of research. *The Journal of the American College of Dentists* 81, 3 (2014), 4.

[36] Department of Homeland Security. 2020. Privacy Impact Assessment Template. Retrieved March 23, 2020 from https://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_template.pdf.

[37] Aaron Yi Ding, Gianluca Limon De Jesus, and Marijn Janssen. 2019. Ethical hacking for boosting IoT vulnerability management: A first look into bug bounty programs and responsible disclosure. In *Proceedings of the 8th International Conference on Telecommunications and Remote Sensing*. 49–55.

[38] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing*.

[39] Serge Egelman, Adrienne Porter Felt, and David Wagner. 2013. Choice architecture and smartphone privacy: There'sa price for that. In *The Economics of Information Security and Privacy*. Böhme R. (Ed.), Springer, Berlin, 211–236.

[40] Malin Eiband, Mohamed Khamis, Emanuel Von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. Understanding shoulder surfing in the wild: Stories from users and observers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4254–4265.

[41] Dean Eckles, Eytan Bakshy, and Michael Bernstein. 2014. Big Experiments: Big Data's Friend for Making Decisions | Facebook. Retrieved March 02, 2020 from https://www.facebook.com/notes/facebook-data-science/big-experiments-big-datas-friend-for-making-decisions/10152160441298859/.

[42] Adrienne Porter Felt, Serge Egelman, and David Wagner. 2012. I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns. In *Proceedings of the 2nd ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*. ACM, New York, NY, 33–44. DOI: https://doi.org/10.1145/2381934.2381943

[43] Rachel L. Finn, David Wright, and Michael Friedewald. 2013. Seven types of privacy. In *European Data Protection: Coming of Age*. S. Gutwirth, R. Leenes, P. de Hert, and Y. Poullet (Eds.), Springer, 3–32.

[44] Nuno Fortes, Paulo Rita, and Margherita Pagani. 2017. The effects of privacy concerns, perceived risk and trust on online purchasing behaviour. *International Journal of Internet Marketing and Advertising* 11, 4 (2017), 307–329.

[45] Sheera Frenkel and Kate Conger. 2018. Facebook's Security Chief to Depart for Stanford University—The New York Times. Retrieved January 15, 2019 from https://www.nytimes.com/2018/08/01/technology/facebook-security-alex-stamos.html.

[46] Kevin Granville. 2018. Facebook and Cambridge analytica: What you need to know as fallout widens. *The New York Times*. Retrieved July 07, 2020 https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

[47] Frances S. Grodzinsky and Herman T. Tavani. 2011. Privacy in the cloud: Applying Nissenbaum's theory of contextual integrity. *ACM SIGCAS Computers and Society* 41, 1 (2011), 38–47.

[48] Jens Grossklags and Alessandro Acquisti. 2007. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *Proceedings of 6th Workshop on the Economics of Information Security*.
[49] The Guardian. 2018. Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach | News. Retrieved September 26, 2018 from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.
[50] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. 2013. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 153–162.
[51] Drew Harwell. 2019. The Pregnancy-Tracking App Ovia Lets Women Record Their Most Sensitive Data for Themselves—and Their Boss—The Washington Post. Retrieved February 04, 2020 from https://www.washingtonpost.com/technology/2019/04/10/tracking-your-pregnancy-an-app-may-be-more-public-than-you-think/?arc404=true.
[52] Miguel Helft. 2010. Google Alters Buzz Service over Privacy Concerns —The New York Times. Retrieved January 9, 2019 from https://www.nytimes.com/2010/02/15/technology/internet/15google.html.
[53] Kashmir Hill. 2012. How Target Figured Out a Teen Girl Was Pregnant before Her Father Did. Retrieved August 27, 2019 from https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.
[54] Kashmir Hill. 2014. OkCupid Lied to Users about Their Compatibility as an Experiment. Retrieved July 6, 2018 from https://www.forbes.com/sites/kashmirhill/2014/07/28/okcupid-experiment-compatibility-deception/.
[55] Laura M. Holson. 2009. Putting a Bolder Face on Google—The New York Times. Retrieved January 30, 2019 https://www.nytimes.com/2009/03/01/business/01marissa.html?pagewanted=3&mtrref=undefined&gwh=2970AE389901E20D8913B8FB7ABB5DBE&gwt=pay.
[56] Jason I. Hong, Jennifer D. Ng, Scott Lederer, and James A. Landay. 2004. Privacy risk models for designing privacy-sensitive ubiquitous computing systems. In *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*. ACM, 91–100.
[57] Giovanni Iachello and Jason Hong. 2007. End-user privacy in human–computer interaction. *Foundations and Trends® in Human–Computer Interaction* 1, 1 (2007), 1–137.
[58] Qatrunnada Ismail, Tousif Ahmed, Apu Kapadia, and Michael K. Reiter. 2015. Crowdsourced exploration of security configurations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 467–476.
[59] Molly Jackman and Lauri Kanerva. 2016. Evolving the IRB: Building robust review for industry research. *Washington and Lee Law Review Online* 72, 3 (2016), 442.
[60] Yunhan Jack Jia, Qi Alfred Chen, Shiqi Wang, Amir Rahmati, Earlence Fernandes, Zhuoqing Morley Mao, Atul Prakash, and Shanghai JiaoTong Unviersity. 2017. ContexIoT: Towards providing contextual integrity to appified IoT platforms.In *Proceedings of the 24th Annual Network and Distributed System Security Symposium*.
[61] Haojian Jin, Tetsuya Sakai, and Koji Yatani. 2014. ReviewCollage: A mobile interface for direct comparison using online reviews. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services*. 349–358.
[62] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 453–456.
[63] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. 43–52.
[64] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1168–1176.
[65] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 24 (2014), 8788–8790.
[66] Katharina Krombholz, Adrian Dabrowski, Matthew Smith, and Edgar Weippl. 2017. Exploring design directions for wearable privacy. In *Proceedings of USEC Mini Conference*. Internet Society. https://www.internetsociety.org/doc/exploring-design-directions-wearable-privacy.
[67] Ponnurangam Kumaraguru and Lorrie Faith Cranor. 2005. *Privacy Indexes: A Survey of Westin's Studies* (CMU-ISRI-5-138), Technical Report. Institute for Software Research International, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
[68] Michelle Kwasny, Kelly Caine, Wendy A. Rogers, and Arthur D. Fisk. 2008. Privacy and technology: Folk definitions and perspectives. In *Proceedings of the CHI'08 Extended Abstracts on Human Factors in Computing Systems*. 3291–3296.
[69] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. 23–34.

[70] Haein Lee, Hyejin Park, and Jinwoo Kim. 2013. Why do people share their context information on Social Network Services? A qualitative study and an experimental study on users' behavior of balancing perceived benefit and risk. *International Journal of Human-Computer Studies* 71, 9 (2013), 862–877.
[71] Sam Levin. 2017. Facebook Told Advertisers It Can Identify Teens Feeling 'insecure' and 'worthless' | Technology | The Guardian. Retrieved February 17, 2020 from https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens.
[72] Yang Li, Jason I. Hong, and James A. Landay. 2007. Design challenges and principles for Wizard of Oz testing of location-enhanced applications. *IEEE Pervasive Computing* 6, 2 (2007), 70–75.
[73] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I. Hong. 2014. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Proceedings of the 10th Symposium on Usable Privacy and Security*.
[74] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49, 2 (2017), 433–442.
[75] Di Liu, Randolph G. Bias, Matthew Lease, and Rebecca Kuipers. 2012. Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–10.
[76] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. 61–70.
[77] Lucidchart. 2018. Context Data Flow Diagram Template. Retrieved December 22, 2018 from https://www.lucidchart.com/pages/templates/data-flow-diagram/context-data-flow-diagram-template.
[78] Miguel Malheiros, Sören Preibusch, and M. Angela Sasse. 2013. "Fairly truthful": The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure. In *Proceedings of the International Conference on Trust and Trustworthy Computing*. Springer, 250–266.
[79] Scott McCloud. 1993. *Understanding Comics: The Invisible Art*. Northampton, Mass. William Morrow Paperbacks.
[80] Aleecia M. McDonald and Lorrie Faith Cranor. 2010. Americans' attitudes about internet behavioral advertising practices. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*. ACM, New York, NY, 63–72. DOI: https://doi.org/10.1145/1866919.1866929
[81] Matt McFarland. 2018. Amazon Go: No Cashiers, Hundreds of Cameras, and Lots of data—CNN. Retrieved February 4, 2020 from https://www.cnn.com/2018/10/03/tech/amazon-go/index.html.
[82] Alan McQuinn and Daniel Castro. 2019. *The Costs of an Unnecessarily Stringent Federal Data Privacy Law*. Technical Report. Information Technology and Innovation Foundation, Washington, D.C.
[83] MeetingKing. 2020. Calculate Meeting Cost. Retrieved March 2, 2020 from https://meetingking.com/meeting-cost-calculator/.
[84] Rolf Molich and Jakob Nielsen. 1990. Improving a human-computer dialogue. *Communications of the ACM* 33, 3 (March 1990), 338–348. DOI: https://doi.org/10.1145/77481.77486
[85] Vivian Genaro Motti and Kelly Caine. 2015. Users' privacy concerns about wearables. In *Proceedings of the International Conference on Financial Cryptography and Data Security*. Springer, 231–244.
[86] Stuart Myerscough, Ben Lowe, and Frank Alpert. 2008. Willingness to provide personal information online: The role of perceived privacy risk, privacy statements and brand strength. *Journal of Website Promotion* 2, 1–2 (2008), 115–140.
[87] Arvind Narayanan and Bendert Zevenbergen. 2015. No encore for encore? Ethical questions for web-based censorship measurement. Retrieved September 24, 2015 from https://ssrn.com/abstract=2665148 or http://dx.doi.org/10.2139/ssrn.2665148
[88] Lily Hay Newman. 2018. The Privacy Battle to Save Google from Itself | WIRED. Retrieved January 10, 2019 from https://www.wired.com/story/google-privacy-data/.
[89] Jakob Nielsen. 1995. Applying discount usability engineering. *IEEE Software* 12, 1 (1995), 98–100.
[90] Jakob Nielsen. 1995. Heuristic Evaluation: How-To: Article by Jakob Nielsen. Retrieved January 04, 2019 from https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/.
[91] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 249–256. DOI: https://doi.org/10.1145/97243.97281
[92] Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.
[93] National Institute of Standards and Technology. 2015. Draft NISTIR 8062: An Introduction to Privacy Engineering and Risk Management in Federal Systems. Retrieved from https://csrc.nist.gov/csrc/media/publications/nistir/8062/draft/documents/nistir_8062_draft.pdf.
[94] National Institute of Standards and Technology. 2017. NISTIR 8062: An Introduction to Privacy Engineering and Risk Management in Federal Systems. Retrieved from https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf.
[95] Information Commissioner's Office. 2020. Data Protection Impact Assessments (DPIAs) | ICO. Retrieved December 30, 2020 from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/.

[96] Cathy O'Neil. 2019. China Knows How to Take Away Your Health Insurance— Bloomberg. Retrieved February 04, 2020 from https://www.bloomberg.com/opinion/articles/2019-06-14/china-knows-how-to-take-away-your-health-insurance.
[97] Stacey A. Page and Jeffrey Nyeboer. 2017. Improving the process of research ethics review. *Research Integrity and Peer Review* 2, 1 (2017), 1–7.
[98] R. Michael Varney Pamela S. Hrubey. 2020. Privacy and Data Protection: Internal Audit's Role in Establishing a Resilient Framework. Retrieved December 28, 2020 from https://www.crowe.com/-/media/Crowe/LLP/folio-pdf-hidden/Privacy_and_Data_Protection_Crowe_IIA_IAF_Joint_Report_CC2015-006.pdf?la=en-US&modified=20200407161139&hash=4F7360FDC3C61D820622DA34FC448C5B1E6F7877.
[99] Clare Payne. 2016. Bad People, or Bad Decisions? | Investment Magazine. Retrieved April 26, 2020 from https://www.investmentmagazine.com.au/2016/11/bad-people-or-bad-decisions/.
[100] Dan Pearson. 2016. Riot Uses LoL Chatlogs to Weed Out Toxic Employees | GamesIndustry.biz. Retrieved February 4, 2020 from https://www.gamesindustry.biz/articles/2016-06-10-riot-uses-lol-chatlogs-to-weed-out-toxic-employees.
[101] Angelisa C. Plane, Elissa M. Redmiles, Michelle L. Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *Proceedings of the 26th USENIX Security Symposium*.
[102] CMU PrivacyGrade. 2015. Grading the Privacy of Smartphone Apps. http://privacygrade.org/.
[103] Leena Rao. 2011. Sexual Activity Tracked by Fitbit Shows Up in Google Search Results | TechCrunch. Retrieved February 4, 2020 from https://techcrunch.com/2011/07/03/sexual-activity-tracked-by-fitbit-shows-up-in-google-search-results/.
[104] Michael Reilly. 2017. Is Facebook Targeting Ads at Sad Teens?—MIT Technology Review. Retrieved February 17, 2020 from https://www.technologyreview.com/s/604307/is-facebook-targeting-ads-at-sad-teens/.
[105] Eric Reis. 2011. *The Lean Startup*, Vol. 27. Crown Business, New York, NY.
[106] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 521–530.
[107] Mary Beth Rosson, John M. Carroll, and Natalie Hill. 2002. *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*. Morgan Kaufmann.
[108] Spencer Rothwell, Steele Carter, Ahmad Elshenawy, and Daniela Braga. 2016. Job complexity and user attention in crowdsourcing microtasks. In *Proceedings of the 3rd AAAI Conference on Human Computation and Crowdsourcing*.
[109] Stuart Schechter and Cristian Bravo-Lillo. 2014. *Ethical-Response Survey Report: Fall 2014*. Technical Report. Technical Report MSR-TR-2014-140. https://www.microsoft.com/en-us/research/publication/ethical-response-survey-report-fall-2014/.
[110] Stuart Schechter and Cristian Bravo-Lillo. 2014. Using ethical-response surveys to identify sources of disapproval and concern with Facebook's emotional contagion experiment and other controversial studies. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/CURRENT20DRAFT20-20Ethical-Response20Survey.pdf.
[111] NOAM SCHEIBER. 2017. How Uber Uses Psychological Tricks to Push its Drivers' Buttons. Retrieved July 8, 2018 from https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html.
[112] Kim Bartel Sheehan. 2002. Toward a typology of Internet users and online privacy concerns. *The Information Society* 18, 1 (2002), 21–32.
[113] Michael Skirpan and Tom Yeh. 2017. Designing a moral compass for the future of computer vision using speculative analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1368–1377.
[114] Smartdraw. 2018. Data Flow Diagram—Everything You Need to Know About DFD. Retrieved December 22, 2018 from https://www.smartdraw.com/data-flow-diagram/.
[115] Daniel J. Solove. 2006. A taxonomy of privacy. *University of Pennsylvania Law Review* 154, 3 (2006), 477.
[116] Sarah Spiekermann. 2012. The challenges of privacy by design. *Communications of the ACM* 55, 7 (2012), 38–40.
[117] Adam Tanner. 2014. *What Stays in Vegas: The World of Personal Data—Lifeblood of Big Business—and the End of Privacy as We Know It*. PublicAffairs.
[118] Gwen Thomas. 2006. *The DGI Data Governance Framework*, Vol. 20. *The Data Governance Institute, Orlando, FL*.
[119] ThoughtWorks. 2015. Security Sandwich | Technology Radar | ThoughtWorks. Retrieved April 25, 2020 from https://www.thoughtworks.com/radar/techniques/security-sandwich.
[120] Tripadvisor. 2019. Number of Characters for Review—Tripadvisor Support Forum. Retrieved February 17, 2020 from https://www.tripadvisor.com/ShowTopic-g1-i12105-k12375287-Number_of_characters_for_review-Tripadvisor_Support.html.
[121] Khai N. Truong, Gillian R. Hayes, and Gregory D. Abowd. 2006. Storyboarding: An empirical determination of best practices and effective guidelines. In *Proceedings of the 6th Conference on Designing Interactive Systems*. ACM, 12–21.

[122] Khai N. Truong, Elaine M. Huang, Molly M. Stevens, and Gregory D. Abowd. 2004. How do users think about ubiquitous computing? In *Proceedings of the CHI'04 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1317–1320.

[123] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proceedings of the 8th Symposium on Usable Privacy and Security*. ACM, New York, NY, Article 4, 15 pages. DOI: https://doi.org/10.1145/2335356.2335362

[124] Lawrence G. Votta Jr. 1993. Does every inspection need a meeting? In *Proceedings of the 1st ACM SIGSOFT Symposium on Foundations of Software Engineering*. 107–114.

[125] Vox. 2018. The Facebook and Cambridge Analytica scandal, explained with a simple diagram. Retrieved September 26, 2018 from https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram.

[126] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *Proceedings of the 6th (2019) ACM Conference on Learning@ Scale*. 1–10.

[127] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I regretted the minute I pressed share" a qualitative study of regrets on Facebook. In *Proceedings of the 7th Symposium on Usable Privacy and Security*. 1–16.

[128] Yang Wang, Gregory Norice, and Lorrie Faith Cranor. 2011. Who is concerned about what? A study of American, Chinese and Indian users' privacy concerns on social network sites. In *Proceedings of the 4th International Conference on Trust and Trustworthy Computing*. Springer, 146–153.

[129] Zhou Wei. 2019. What Your Face May Tell Lenders about Whether You're Creditworthy—WSJ. Retrieved February 4, 2020 from https://www.wsj.com/articles/what-your-face-may-tell-lenders-about-whether-youre-creditworthy-11560218700.

[130] Western City. 2012. The "Front Page" Test: An Easy Ethics Standard—Western City Magazine. Retrieved April 25, 2020 from https://www.westerncity.com/article/front-page-test-easy-ethics-standard.

[131] Wikipedia. 2020. Discounted Cumulative Gain—Wikipedia. Retrieved March 3, 2020 from https://en.wikipedia.org/wiki/Discounted_cumulative_gain.

[132] Wikipedia. 2020. Separation of Duties. Retrieved April 23, 2020 from https://en.wikipedia.org/wiki/Separation_of_duties.

[133] Jenifer S. Winter. 2012. Privacy and the emerging internet of things: Using the framework of contextual integrity to inform policy. In *Proceedings of the Pacific Telecommunication Council Conference*, Vol. 2012.

[134] Molly Wood. 2014. OKCupid Plays With Love in User Experiments—The New York Times. Retrieved July 6, 2018 from https://www.nytimes.com/2014/07/29/technology/okcupid-publishes-findings-of-user-experiments.html?ref=technology&_r=1.

[135] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Lauren Schmidt, Laura Brandimarte, and Alessandro Acquisti. 2014. Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences. In *Proceedings of the 10th USENIX Conference on Usable Privacy and Security*, Vol. 5, USENIX Association, 1–18.

[136] David Wright and Paul De Hert. 2011. *Privacy Impact Assessment*. Vol. 6. Springer Science & Business Media.

Received April 2020; Revised February 2021; Accepted April 2021