Demand for Privacy from Data Brokers: Pre-Analysis Plan*

Joy Wu Avinash Collis Ananya Sen

February 29, 2024

Contents

*Wu: Cornell University, SC Johnson College of Business, zw369@cornell.edu. Collis: Heinz College of Information Systems and Public Policy, Carnegie Mellon University, avinashcollis@cmu.edu; Sen: Heinz College of Information Systems and Public Policy, Carnegie Mellon University, ananyase@andrew.cmu.edu;

1 Introduction

There is a large and pervasive market for personal data: information about peoples' attitudes, behaviors, and attributes. Generated through users' online activities or digitized from various administrative records, these data are collected and curated as assets to be traded and commercialized in data markets. Personal data are often utilized as valuable inputs into algorithms that generate revenue for firms. Prominent players in these data markets are data brokers, who sell access to and usage of the personal data they harvest. Many people are not only inactive in these economic activities, but they may also be unaware of the existence and extent of data brokers' activities. Currently, little is known about people's valuation, perceptions, and demand for data privacy from data brokers.

Our main, empirical task is to test how the availability of users' data to data brokers shapes individuals' demand for data excludability rights. We capture demand for data excludability through the willingness to pay for a real privacy commodity: data deletion from data brokers. We will explore how the demand for privacy from data broker activities may differ depending on whether individuals are aware of data broker activities, the downstream consequences of commercial activity by data brokers, or from which entities brokers source individuals' data.

2 Research Strategy

We will use a survey experiment to capture the revealed preferences and stated attitudes of a representative U.S. population. We plan to use education interventions about how data are collected by brokers and what are the consequences of those collection efforts. In observing respondents' willingness to pay for a data deletion service, we will capture whether individuals demand privacy from data brokers. We plan to recruit 4,000 participants from a representative sample of U.S. internet users in collaboration with YouGov.[1]

In additon to examining how peoples' willingness to pay for privacy respond to interventions about data broker activities, we will explore secondary outcomes and mechanisms such as individuals' beliefs about data exposure, their psychological ownership over their personal data harvested by brokers, and notions of fairness and digital resignation about the personal data market. Given the representative U.S. sample we study, we will document heterogeneity in privacy preferences and beliefs based on demographic characteristics recorded by YouGov as well as those elicited in the survey.

3 Design

We plan to elicit real decisions to pay for personal data monitoring and deletion for a period of one year from 90 major U.S. data brokers (see Appendix Figure A.1). This data deletion service will

[1] YouGov America Inc. (https://business.yougov.com/).

be a real product utilized in our study in collaboration with HelloPrivacy, a data privacy company that offers a subscription service for monitoring and deleting individuals' data acquired by these brokers (see Appendix Figure A.2).

3.1 Pre-Intervention Survey Questions

Pre-intervention survey questions occur before participants are randomized into the intervention groups about data brokers' data sources, the secondary uses of those data, and the impacts of these uses on individuals. These include informed consent to participate in the study, a lottery number selection question, a basic survey about the participant's prior activities that are typically digitized and harvested by data brokers, and a basic introduction to the study with definitions for "personal data" and "data brokers."

After obtaining their informed consent to participate in the study and before any information is provided about the content of the study, we ask respondents to enter into a $50 lottery with a 1-in-50 chance of winning. At this point, respondents are provided again with an active choice to opt-out of the study; however, respondents are allowed to voluntarily end their survey participation at any point. Respondents select a lottery number (out of 50 provided on their screen) and are told, should they win, a final bonus would be provided to them after the study. In addition, they are told that they will be asked to make decisions about whether and how to spend some of these bonus earnings (should they win). After the study, the procedure to select bonus earnings is to use a computer program to randomly select a winning lottery number for *each* respondent, and if the respondent's chosen lottery number matches their winning number, they will earn a $50 bonus (delivered to them as bonus payments by YouGov).[2]

Following the lottery, the pre-intervention background questions ask individuals to indicate their data-sharing and data-deletion experiences. Prior to any information about the study (i.e., that it relates to personal data or data brokers), we ask individuals to identify activities they have conducted that generate data that are typically harvested by data brokers. These questions include whether they filled out a warranty card, created an account on a website using their personal information, registered a change of address with the U.S. Postal Service, and paid for products using a credit card under their name. Second, panelists are introduced to the basic purpose of the study, and basic definitions are provided about "personal data" and "data brokers." On this information page, respondents are asked whether they have previous experience deleting data from data brokers.

3.2 Interventions

Respondents will be randomized into information interventions about how data brokers collect and commercialize data. We designed these interventions based on the May 2014 report, "Data Brokers:

[2] We receive the respondent's survey identification number, but we do not have access to their personal identifiers. We provide YouGov information about which survey responses are receiving bonus payments.

A Call for Transparency and Accountability," from the U.S. Federal Trade Commission (Ramirez et al., 2014, hereafter the "2014 FTC Report"). Information treatments are delivered in the form of an infographic on data collection that maps each stage of the originator of the personal data (i.e., the individual) to how that data is obtained and used by a representative data broker. First, the intervention includes a sample of various and common actions and behaviors of individuals. Second, it illustrates how these activities generate personal data for parties the individuals directly interact with (i.e., second parties). Finally, the intervention describes how those data are traded and collected by a thirty-party data broker—both with those second parties and with other data brokers. Following the investigation of the 2014 FTC Report, we categorize the second parties as two primary "sources" of data for brokers: government and commercial entities (see Appendix Figures A.3 and A.4).

Table 1: Information Treatment Groups

Group	Description	Estimated Sample Size
Control	No information provisions	1,300
Government	Information provision about broker data from *government sources*, what products brokers sell, and the consequences of those products on idnviduals	1,300
Commercial	Information provision about broker data from *commercial sources*, what products brokers sell, and the consequences of those products on individuals	1,300

Notes: Estimated total sample size of 4,000 individuals.

For each treatment group, we also provide identical information on the consequences of data collection by data brokers: what products they sell after using the data they harvest and how those products impact individuals. Following the 2014 FTC Report, we present three categories of commercial products produced by data brokers in the secondary data market: marketing, risk management, and people search.

3.3 Post-Intervention Outcomes

3.3.1 Data Exposure Beliefs

Respondents will be asked to provide predictions about their degree of exposure to the list of 90 data brokers, including both the number and type of data exposed. We elicit predictions after information interventions and before eliciting their willingness to pay for data deletion. These outcomes are the components of measuring beliefs about data availability in our study and testing whether these predictions respond to our interventions. In addition, we utilize these predictions to understand variation in treatment effects on the data privacy valuations (see Section 3.3.2).

3.3.2 Data Privacy Valuations

To elicit revealed preferences, the study first provides all subjects with a lottery to win a $50 bonus at the start of the survey (see Section 3.1). This ensures that the study never forces subjects to pay out of their own pocket (or lose money) during the study. This also ensures that incentive-compatible decisions are elicited from the subject|i.e., they decide the maximum they are willing to spend out of their potential bonus. Subjects do not know if they have won bonus earnings throughout the survey, and it is only revealed to them upon delivery of their final payments after study completion.

The maximum willingness-to-pay prices are elicited using a multiple price list with a first-stage selection decision and a second-stage pricing decision. The willingness of the participant to spend any of their bonus earnings is the first choice. At this stage, they are told that, should they be willing to spend some amount of their bonus, they will also decide the price they are willing to pay (i.e., they are not committing to pay anything until they see the available prices). Then, conditional on being willing to pay, they are asked to choose whether they would be willing to pay or not pay prices ranging from $5 to $50 in increments of $5.

Explanations are provided in both decision stages that it is in their best interest to answer honestly. In addition, the price decision screen explains that any one of the prices can be the final price, in which the respondent's decision for that price would be implemented for real, should they win the bonus. Following the Becker et al. (1964) method, the final and unknown price is randomly selected. We will use a computer program to randomly select a final price for each subject who wins the lottery, and we will implement their chosen accept or reject decision for that price.

Finally, at this stage in the survey, we also ask respondents to answer in an open-text some of their reasonings behind their decisions thus far.

3.3.3 Stated Attitudes

At the last stage of the survey, all stated attitudes questions are presented as a list of statements, in which respondents are asked to answer their agreement or disagreement on a Likert scale (1 = Strongly Disagree . . . 5 = Strongly Agree). Each category of these attitudes is randomized in the order presented to each respondent. In addition, within each category, the statements are randomly ordered.

1. Psychological Ownership of Personal Data. Respondents are asked the degree to which they feel psychological ownership over the data that brokers have harvested about them.[3]
   - \I feel like data about me that have been collected by data brokers are mine."
   - \I feel a very high degree of personal ownership of the data that brokers have collected about me."

[3] We use items adapted from Peck and Shu (2009).

- ˆ \I feel like I own the data the brokers have collected about me."
- ˆ Attention check: \Select `Strongly disagree' if you are reading this."

2. Attitudes About Data Brokers. These items capture respondents' opinions about data brokers' activities.
   - ˆ \It's unfair that individuals have to pay to delete their data from brokers."
   - ˆ \If I delete my data from brokers, they will just collect my data again."
   - ˆ \Data brokers should not have the right to collect personal data about individuals."
   - ˆ \Data brokers should not be allowed to sell personal data to third parties."
3. Data-Sharing Activities. This set of items capture respondents' everyday activities that involve data-sharing.
   - ˆ \I frequently comment, post, like, and re-share things on social media platforms, such as Facebook, Instagram, or TikTok, non-anonymously (e.g., under a real name or alias that can easily identify me)."
   - ˆ \I often subscribe to frequent buyer programs, ll-out warranty cards, or sign-up for sweepstakes using my email, personal information, or address."
   - ˆ \I often have location-tracking features on my mobile phone turned on while doing activities such as shopping in physical stores or dining out at restaurants."
   - ˆ \My friends or relatives post a lot of personal information about themselves publicly online, some of which contain information about me."
4. Opinions About Policies. These items capture their stated attitudes toward privacy policies, privacy rights, and political leanings.
   - ˆ \I tend to side with Republicans on most issues."
   - ˆ \I tend to side with Democrats on most issues."
   - ˆ \I trust that the free market leads to appropriate privacy protection."
   - ˆ \I believe that privacy is a fundamental human right."

3.4 Survey Chronology

A summary of the survey chronology is presented in Figure 1.

Figure 1: Survey Chronology

3.5 Non-Survey Data

Additional data on panelists are provided by YouGov, which are not asked in our survey. These include:

Birth year	2016 U.S. presidential vote
Gender	2020 U.S. presidential vote
Race	State of residence
Education	Voter registration status
Marital status	Political ideology
Employment status	Church attendance
Family income	Religion
Number of children under 18	Born again
Political party	Importance of religion
Political party leaning	Frequency of prayer

3.6 Pilots

We conducted three pilots (approximately 300 participants each) of the control group condition with a representative sample of Google Cloud Research panelists. These pilots allowed us to gather an approximate understanding of the prices individuals are willing to pay for the data deletion service, increasing our price list maximum from $25 to $50. In a fourth pilot, we examined whether randomizing the psychological ownership Likert-type questions before or after the elicitation of privacy valuations interacted with responses. We found no conclusive evidence of this behavior.

We provided a final draft of the survey in Qualtrics to YouGov's survey team to program the final version in their custom interface. A feature of the YouGov survey not included in our Qualtrics draft is the ability to auto-fill prices in the price list to prevent price-switching responses. A soft launch was conducted with 100 YouGov panelists to check for any bugs in the programming, as well as to conduct a trial run of our process for selecting lottery winners. We conducted a final pilot of

300 respondents with Cloud Research to evaluate comprehension of instructions. No substantive changes to our survey were made after these last two pilots, except to correct typos in Likert-type items that used the term \consumers" to instead use \individuals" in maintaining consistency with the terminology of the infographics.

4 Framework

Our study concerns the valuation of a privacy-preserving product in the presence of uncertainty about data exposure and inattention towards data-broker activities. We assume this valuation to be driven by the individual's exposure to data brokers, e, and their preferences for the privacy they would receive from data protection from brokers' access, $v(e)$, which includes both intrinsic and instrumental values for privacy (Lin, 2022, Farrell, 2012). Consistent with the well-documented information disadvantage that individuals have about their valuation of privacy, the individual perceives the value of the privacy product to be $\hat{v} = v(\hat{e})$. We allow $v(\cdot)$ and $\hat{e}$ to vary by experimental group $\in \{0; G; C\}$ (respectively, the Control, Government, and Commercial groups), which are designed to exogenously shift beliefs about e and attention towards the components that determine $v(\cdot)$.

Individuals may change their estimates $\hat{e}$ after learning new information about the sources of data harvested by brokers. For example, if they learn that one source of data comes from registration websites of online retailers, and the individual often makes a lot of these online retailer accounts, they may increase their estimated exposure to data brokers.

Individuals may also care yet be inattentive towards broker activities, downstream consequences of these activities, or the entities brokers are sourcing from, which make certain components contained in $v(\cdot)$ opaque without salient information.[4] Individuals might have di erent preferences for the privacy they enjoy by preventing these data utilization activities and the manner in which brokers obtained these data. Downstream consequences, as provided in our information provisions, include monetization activities of brokers (e.g., pro ting from marketing or people search products) and instrumental impacts (e.g., online stalking, price discrimination, personalized ads). The source of the data, holding xed the downstream activities and consequences, can be whether the data brokers obtain personal data from government or commercial entities.

[4] We simplify this concept of inattention by allowing $v(\cdot)$ to change based on environmental signals, but we assume that the true preferences of individuals are more stable and can be decomposed into separate components, such as protection from brokers harvesting commercial data, data monetization by brokers, price discrimination, etc., each with their own inattention parameters that vary based on the salience of a signal about that component.

5 Hypotheses and Empirical Strategy

5.1 Main Outcomes and Hypotheses

Our main hypotheses concern the revealed preference elicited in the study: the maximum price individuals are willing to pay (y) for the privacy commodity that deletes their personal data from 90 major U.S. data brokers. We expect that this value is in uenced by the individual's predicted data exposure and the awareness of data broker sources, their commercial activities, and the consequences of broker activities. We test the null:

$$y\ (\hat{e}_G; v_G) = y\ (\hat{e}_0; v_0) \qquad \text{(H1-a)}$$

$$y\ (\hat{e}_C; v_C) = y\ (\hat{e}_0; v_0) \qquad \text{(H1-b)}$$

In addition, comparing Government with Commercial information interventions allows us to isolate how the source of data di erentially impacts data privacy valuations, holding salient and xed the downstream activities and consequences. We test the null:

$$y\ (\hat{e}_G; v_G) = y\ (\hat{e}_C; v_C) \qquad \text{(H2)}$$

5.2 Secondary Outcomes and Hypotheses

To explore mechanisms, we will examine several stated beliefs and attitudes outcomes of the experimental interventions.

First, we will measure whether the average level of predicted exposure di ers between treatments. We will examine these in the number of brokers (out of 90) the individual predicts have harvested her data. We will also examine these in the number of categories of data (e.g., age, names, addresses) predicted by the individual and the propensity to select any of the categories. We test the null:

$$\hat{e}_G = \hat{e}_0 \qquad \text{(H3-a)}$$

$$\hat{e}_C = \hat{e}_0 \qquad \text{(H3-b)}$$

$$\hat{e}_C = \hat{e}_G \qquad \text{(H4)}$$

Finally, we will also explore the slope of privacy valuations relative to data exposure. As a baseline, we will explore an understanding of the value of privacy relative to perceived data exposure, in the number of predicted data brokers (out of 90) that have collected personal information about the

individual: [5]

$$\frac{@y(\hat{e}; v)}{@\hat{e}} = 0 \qquad \text{(H5)}$$

In addition, we can explore heterogeity in treatment e ects based on these beliefs. For instance, we will assess whether valuations for privacy corresponds to changes in predicted exposure the same across all environments.

Finally, to further explore mechanisms underlying potential di erences in revealed valuations due to the experiment interventions, we also test the impact of interventions on the stated attitudes[6] towards psychological ownership,[7] fairness attitudes,[8] and control.[9]

5.3 Heterogeneity Analyses

We will examine heterogeneity in privacy valuation and treatment e ects based on the pre-intervention survey items (see Section 3.1) and the post-intervention Likert-type attitudes (see Section 3.3.3). We will also examine non-survey demographic variables provided by YouGov (see Section 3.5).

5.4 Sample Exclusion Criteria

The exclusion criteria for workers are programmed into the survey by the third-party YouGov and only completed responses are considered and delivered to us. YouGov also conducts standard data removal procedures before the nal delivery of the data to our team. These broadly include the removal of respondents who are speeding through the survey, excessively skipping survey questions,[10] excessively \straight-lining" grids, [11] and suspected of fraudulent or bot-like activity.[12]

While we collect data on comprehension question attempts (related to the information interventions) and one attention check in the survey, we do not intend to exclude respondents based on these survey items. However, we intend to use these to measure the quality of our sample and conduct robustness checks. In general, comprehension questions are utilized in our survey to better

[5] We consider the possibility of this relationship to be in either direction or even non-linear. In other survey items, we survey the role of digital resignation, which may interact with the predicted exposure of individuals.

[6] See items in Section 3.3.3.

[7] The feeling of perceived ownership of data about the individual collected by brokers.

[8] Whether it is fair that individuals must pay to delete their data from brokers.

[9] This relates to the \digital resignation" among respondents (Draper and Turow, 2019).

[10] We require all survey responses to be completed before continuing to the next survey item and completing the survey. Thus, this deletion criteria does not apply to our study.

[11] We have avoided questions formatted in large grids. For example, Likert-type items are presented one at a time. The only exceptions are the multiple price list (which is programmed with assistance to prevent price-switching behavior by automatically populating non-acceptance|and prevents acceptance|of higher prices after respondent selects a price they are unwilling to pay) and the few pre-intervention background questions (with only three answer options each).

[12] These include activities such as cross-referencing the responses provided in the survey to those in YouGov's internal database, monitoring survey meta-data (e.g., geolocation and time stamps), and reviewing response patterns across open-ended questions and attention checks.

support respondents in learning the information provided in the survey.[13]

5.5 Empirical Specifications

Our empirical tests are based on a series of regression specifications. Our main outcome variable is the maximum willingness to pay choice for the data deletion service. Following the style of our valuation elicitation, we can split this into two stages. First, the willingness to pay any price

$$\text{Pay}_i = \begin{cases} 1; & y_i > 0; \\ 0; & y_i \quad 0; \end{cases}$$

where y_i is the true, unobserved maximum willingness-to-pay price for individual i. Second, for those who are willing to pay, the observed maximum price choice when $\text{Pay}_i = 1$ is

$$\text{Price}_i = \begin{cases} 0; & 0 < y_i < 5; \\ 5; & 5 \quad y_i < 10; \\ \vdots & \vdots \\ 45; & 45 \quad y_i < 50; \\ 50; & 50 \quad y_i : \end{cases}$$

In addition, we can combine the two stages of the valuation choice with stronger assumptions, allowing $\text{Pay}_i = \text{Price}_i = 0$ to indicate $y_i < 5$. For the remainder of this section, we will refer to this revealed price choice as y_i, and note that we can decompose our results for the extensive margin (i.e., Pay_i) and the intensive margin (i.e., Price_i).

Our main specification is:

$$y_i = + \sum \quad _i + "_i$$

where $_i$ is equal to one if the respondent is in group and zero otherwise, and $"_i$ is the error term. This allows us to test H1-a, H1-b, and H2: $\hat{}_0 = \hat{}_C$, $\hat{}_0 = \hat{}_G$, and $\hat{}_C = \hat{}_G$. In our secondary analyses, we will examine the average effects of conditions on $e_i \in [0; : : : ; 90]$, using the specification

$$\hat{e}_i = + \sum \quad _i + "_i$$

to test H3-a, H3-b, and H4: $\hat{}_0 = \hat{}_C$, $\hat{}_0 = \hat{}_G$, and $\hat{}_C = \hat{}_G$. To examine the relationship between

[13] We include three short "True" or "False" questions specific to each information intervention in order to help respondents learn and comprehend the information. For example, one question provides an example commercial or government record (e.g., online retailers or the U.S. postal service) and asks the respondent whether data brokers can collect these data.

predicted exposure and privacy valuations, we specify:

$$y_i = \quad + \quad \sum \quad {}_i(\quad + \quad \hat{e}_i) + \varepsilon_i$$

For hypothesis 5, we can examine whether $\hat{\ }_0, \hat{\ }_G; \hat{\ }_C = 0$. To explore associations between predictions and treatment effects, we can, for instance, use this same specification to test for the equality of the slope coefficient.

We will also examine the impacts of our interventions on stated attitudes with 5-point Likert-scales (1 = Strongly disagree, . . . , 5 = Strongly agree). For the items related to psychological ownership, we will construct two alternative dummy measures: (1) whether values average to a value greater than or equal to 4 and (2) whether the respondent indicated a 4 or a 5 on all three psychological ownership items. Since fairness and control items each are based on one survey item, we will construct the indicator variable of whether these responses are a 4 or 5.

Finally, heterogeneity analyses will be conducted as extensions of these regression specifications.

References

Becker, G., M. Degroot, and J. Marschak (1964). Measuring utility by a single-response sequential method. Behavioral Science 9(3), 226–232.

Draper, N. A. and J. Turow (2019). The corporate cultivation of digital resignation. New Media & Society 21(8), 1824–1839.

Farrell, J. (2012). Can privacy be just another good? Journal of Telecommunication and High Technology Law 10, 251–264.

Lin, T. (2022). Valuing intrinsic and instrumental preferences for privacy. Marketing Science 41, 663–681.

Peck, J. and S. B. Shu (2009). The effect of mere touch on perceived ownership. Journal of Consumer Research 36(3), 434–447.

Ramirez, E., J. Brill, M. K. Ohlhausen, J. D. Wright, and T. McSweeny (2014, May). Data brokers: A call for transparency and accountability. Technical report, United States Federal Trade Commission.

A Appendix

Figure A.1: List of 90 Major U.S. Data Brokers

Notes : Broker list taken from those monitored by HelloPrivacy Inc. The actual number of brokers are taken from their propriety database, after purchasing access to their API, rather than the \50+" advertised to potential customers in Appendix Figure A.2.

Figure A.2: Data Deletion Service

Notes : Screenshot from the commercially advertised version of the data deletion product of HelloPrivacy (https://brandyourself.com/remove-info-data-brokers; accessed February 28, 2023). We collaborate with this company to purchase and utilize this product (i.e., one annual subscription) as the real privacy commodity that participants our study can purchase. We do not disclose the identity of the company or deletion product to our respondents, in order to minimize direct price comparisons that may interact with their responses in our survey.

Figure A.3: Government Data Sources Information Intervention