Bounded Rationality in Selection Markets: Evidence from Consumer Data Sharing *

Mohaddeseh Heydari-Nejad † Verina F. Que ‡

October 2026

Abstract

This paper empirically studies how bounded rationality affects the welfare consequences of information policies in markets with endogenous participation. We examine this question in data markets, where the terms consumers receive depend on the pool of individuals who share their data. Using survey responses from 3,132 Dutch individuals about their trust in how nine types of organizations use personal data, together with one actual data-sharing decision per respondent, we estimate a cognitive-hierarchy model of data disclosure. We find that seventy-three percent of consumers are naive: they fail to account for the selection of individuals into data sharing. In a counterfactual that raises the cognitive sophistication, decision losses fall by 81 percent, but average consumer welfare falls by 4.3 percent and only 11 percent of consumers gain. The loss arises because personalized terms redistribute surplus among those who share. Its size depends on the joint distribution of consumer types and cognitive sophistication. Welfare rises when more than about half of the differences in personalized terms reflect better matching or other efficiency gains rather than transfers between consumers.

Keywords: Privacy, data markets, bounded rationality, cognitive hierarchy, adverse selection, structural estimation.
JEL codes: D12, D18, D82, D91, L86.

*Acknowledgments: We are deeply grateful to Avi Goldfarb for his continuous guidance and support throughout this project. We thank Anthony Dukes, Garrett Johnson, Jeff Prince, Jidong Zhou for their great feedback. We thank seminar participants at UCLA, and also thank all participants in the weekly meetings of Lab for the Management of Science & Technology (L-MOST) at Rotman. We used AI tools to assist with editing the manuscript. All errors are our own. *The authors did not receive any external funding for this research and declare no financial relationships or other potential conflicts of interest*

†Kelley School of Business, Indiana University. Email: mheydari@iu.edu
‡Nanyang Business School, Nanyang Technological University. Email: verina.que@ntu.edu.sg.

1 Introduction

Regulators increasingly require organizations to make the use of personal data transparent to consumers. For example, in the European Union, traders have been required to disclose personalized prices since 2022. New York has required firms to tell consumers when a price was set by an algorithm using their personal data since 2025, and in August 2026, the Federal Trade Commission proposed to treat undisclosed personalized pricing as a deceptive practice. Such policies do not restrict personalization directly, but they make salient to consumers that the terms they receive differ from the terms others receive. A disclosure therefore leads consumers to ask where they stand relative to others, and so to decide better whether to share their data.

In markets where personalized payoffs depend on the equilibrium composition of data collected by organizations, however, individual sharing choices generate selection externalities. Such policies then shift the composition of the pool, inducing an equilibrium payoff channel that counteracts or amplifies the direct benefits of better-informed individual decision-making. The net welfare consequence in such settings therefore depends on the interaction between individual mistake-correction and equilibrium selection. For example a consumer who joins a loyalty program, consents to tracking, or contributes data to a public program receives personalized terms as well as a common benefit financed by what the program earns from all of its members. Her payoff from sharing therefore depends on how her type compares with other members, as in Akerlof (1970) and Einav et al. (2010).

This paper investigates the welfare and distributional consequences of transparency when personalized terms depend on the composition of the sharing pool. We find that most consumers do not account for their position relative to others, and decide as if they were average members. As a result, they do not sort across organizations, and data pools remain mixed. A policy that makes consumers reason about the pool, improves their decisions, but it also leads them to sort. Higher-type consumers leave the settings that take more from

them and join those that favor them. This sorting is costly for the other sharers. When a consumer leaves a program that takes more from her, the program loses what it earned from her, and the common benefit falls for everyone who stays. When she joins a program that favors her, the better terms she receives are paid for by the other members. Her gain is largely a transfer from other consumers, and because consumers now join or leave programs for these transfers rather than for the value of sharing itself, participation becomes less efficient. When personalized terms mainly redistribute surplus, the losses to others exceed the gains from better decisions and average welfare falls. The consumers who lose the most are those the policy does not inform. They keep making the same decisions but still bear the worse terms in their pools.

Our empirical analysis utilizes panel data from the LISS panel, a representative sample of Dutch households (CentERdata, 2024). In 2024, these respondents were asked to donate the metadata of their WhatsApp messages to academic researchers. Before deciding whether to share these data, they were asked to rate their trust in nine types of organizations with their personal data. These organizations range from the national statistical bureau and academic researchers to app developers, technology companies, and social-media platforms. Two features make the setting suitable for our question. First, the panel structure observes each respondent's trust ratings across organizations with heterogeneous data practices, allowing us to identify organization-specific degree of personalization from within-respondent variation while controlling for individual heterogeneity in trust scale usage and privacy baseline preferences. Second, we observe a real data sharing decision. Whether the respondent donated her data to the researchers at the end of the study. This allows us to anchor the trust ratings to actual data-sharing behavior for each individual, enabling us to map ordinal survey ratings into a structural model of binary participation.

The descriptive evidence shows three patterns. Trust declines from public to commercial organizations. This decline is steeper for respondents who better understand what sharing entails, as measured by a comprehension quiz in the survey. And within person, the decline

depends on the respondent's position in the income distribution, it is steeper for those above the average and flatter for those below it, and this pattern is stronger among respondents with higher education. These patterns suggest that consumers may differ in accounting for where they stand relative to others.

We build a cognitive-hierarchy model following Camerer et al. (2004), and estimate how far consumers account for their position relative to the other sharers, and use the estimates to evaluate the welfare and distributional effect of raising their sophistication. Each organization runs a competitive data program where members who join receive a common benefit and a payoff that depends on their type through an organization-specific targeted treatment. This general model captures type-dependent transfers coming from personalization as well as common transfers from sharing their data. So, we model a member's personalized payoff depends on her type relative to the pool average. Naive consumers do not account for their position and decide as if they were the average member. More sophisticated consumers forecast the pool as a composition of consumers with lower levels of sophistication and place themselves relative to it. We use a cognitive hierarchy because the object of reasoning in this paper is the composition of a pool formed by a heterogeneous population, and the hierarchy delivers each level's belief about that composition.

To build a model that captures behavior across organizations, we take income as the individual's type that an organization learns if she shares her data. Income is the characteristic that we observe for all respondents and is most directly tied to willingness to pay. Our model therefore captures the forms of personalization that rely on what an organization can infer about an individual's willingness to pay for its product or service, such as personalized prices, targeted offers, or differential service or policies[1]. Our model allows the direction of personalization to differ across organizations. Some organizations may take more from higher-income individuals, while others may favor them. We treat an individual's

[1]Organizations may also learn other characteristics from the data consumers share. As a robustness check, we let the type be an index of income, education, age, gender, and digital literacy, with weights estimated jointly with the other parameters. The qualitative conclusions do not change

trust ratings and her actual sharing decision as two measures of the same perceived payoff from sharing her data. The trust ratings tell us how this payoff varies across the nine organizations, and the sharing decision anchors this rating to actual sharing behavior. A naive consumer's trust varies with her type only through her intrinsic privacy preference, which is common to all organizations. However, a more sophisticated consumer's trust also responds to the organization-specific personalization practice. We do not directly observe the realized consequences of consumers' choices. We measure a naive consumer's mistake against the decision of an otherwise similar consumer who accounts for her position. As is standard in the cognitive hierarchy (Camerer et al., 2004; Goldfarb and Xiao, 2011), consumers at all levels have the same payoff function, and they differ only in how they perceive their position in the pool, which a naive consumer ignores and a more sophisticated consumer forecasts from the behavior of lower levels. In our model, a consumer's sophistication level is also independent of her privacy cost given her type.

Our estimates show four main results. First, most consumers do not reason about selection, 73 percent are naive, 23 percent reason one step ahead, and 4 percent two steps. Second, consumers expect organizations to use their data in different ways. Higher-type consumers expect public and research organizations to on average favor them and most commercial organizations, from the internet provider to social media, to take more from them through personalization. Third, bounded rationality limits selection. We define naive subsidy as the average income difference between the pool at the estimated sophistication and the pool under a counterfactual rational expectations scenario. Pools at technology companies and social media contain about 0.45 standard deviations of income more than under rational expectations. Data pools at public and research organizations contain 0.25 to 0.30 standard deviations of income less than under rational expectations. Naive consumers keep higher-income sharers in the commercial pools that selection would push out.

Fourth, raising sophistication has a large decision benefit and a larger equilibrium cost on those who share. We compute welfare from the payoffs consumers actually receive, not

from the payoffs they perceive when they decide (Bernheim and Rangel, 2009)[2]. As we do not observe how organizations actually use the data, we assume that consumers know how organizations differ. For example, that technology companies on average use personal data more intensively than the statistical bureau does. This is plausible, because such differences are widely discussed and are reflected in the clear ranking of trust across organizations in our data. What is harder for consumers is to reason about who else shares. The terms a consumer receives depend on how her type compares with other sharers, and the pool of sharers is not observed when she decides. A naive consumer ignores this endogenous participation in the pool and decides as if she were an average member. This is the mistake that higher levels of sophistication corrects. Raising the share of sophisticated consumers from 27 to 88 percent corrects 81 percent of consumers' errors. The re-sorting of pools lowers average welfare by 4.3 percent of its baseline level. Only 11 percent of consumers gain, all of them initially naive. The gains are concentrated at the tails of the income distribution, where consumers are furthest from the pool average and have the most to correct. Consumers who the policy leaves naive lose the most, because they bear the equilibrium worsening of pools without any decision gain, same as the consumers who were sophisticated before the intervention.

Identification comes from how trust varies with respondents' type across organizations. A naive individual does not consider how her type compares with other sharers, so her type changes her trust by the same amount at every organization, through only her privacy taste. For a sophisticated individual, type above the pool average helps her at organizations that favor higher types and hurts her at organizations that take more from them, so her type changes her trust differently across organizations. The share of respondents whose trust follows this pattern identifies the share of sophisticated consumers, and which organizations they trust more or less identifies how each organization treats higher types. The key assumption is that, apart from this reasoning, a consumer's type affects her trust equally at every

[2]A welfare measure based on perceived rather than realized payoffs would record almost no change, because it evaluates each consumer at the pool she expects.

organization. The donation decision links trust to actual sharing. As in other applications of the cognitive hierarchy (Camerer et al., 2004; Goldfarb and Xiao, 2011), how sophisticated consumers divide between higher levels relies on the structure of the model.

In our baseline model, the data program is break-even, so, the personalization terms are transfers among members, it gives some consumers better terms than the average and others worse. In this environment, the all-naive market is efficient, and sorting has private but no social value (Hirshleifer, 1971). In our baseline, we quantify the size of the two channels, their distribution across consumers, and the point at which the welfare could potentially increase. Following Handel and Kolstad (2015), we investigate an environment that personalization can go beyond pure redistribution. We keep everything at the estimated model and allow a portion of the personalization to deliver real surplus, such as better matching. In our estimates, correcting consumers' reasoning about their relative position raises welfare only if more than 49 percent of what personalization does is create value rather than redistribute it.

The results have several implications for policy. First, consent rates are not a sufficient statistic for the effects of data policy. In a pooled market, a policy can re-sort the pool substantially while the consent rate move less. In our counterfactual, the composition of pools changes substantially. Second, the effects of transparency are distributional. Individuals who the policy does not reach are worse off than before, as they bear the re-sorting of pools without any gain from better decisions. In our sample, naive consumers are concentrated at the bottom of the income distribution, and an information campaign that fails to reach them would harm them the most. Third, the effect of transparency on consumer welfare depends on the welfare relevance of personalization, and when enough of personalization creates real surplus rather than redistributing it, transparency can be welfare-improving.

The paper contributes to several bodies of literature. First, it brings bounded strategic reasoning to the economics of privacy and personalized pricing (Taylor, 2004; Acquisti et al., 2016; Bergemann et al., 2022; Ali et al., 2023). The externality we study falls on those

who continue sharing data, unlike the inference externality borne by those who do not share (Choi et al., 2019; Goldfarb and Que, 2023). Second, it extends field applications of the cognitive hierarchy, which have studied firms (Goldfarb and Xiao, 2011; Hortaçsu et al., 2019), inferences from films withheld from critics (Brown et al., 2012), and consumers in a lottery game (Östling et al., 2011), to data markets, where consumers' personalization payoffs depend on others. Benndorf et al. (2015) show in an experiment that limited reasoning attenuates unraveling, we study this mechanism in data markets. Third, it contributes to work on consumer frictions in selection markets. Handel (2013) shows that reducing inertia in health insurance worsens adverse selection, and Gabaix and Laibson (2006) show that educating naive consumers removes the subsidy they provide. We study the failure to account for the position among the other participants, measure how common it is and study its welfare and distributional consequences in data disclosure setting.

Section 2 reviews the related literature. Section 3 describes the setting and the data, Section 4 the model, and Section 5 identification and estimation. Section 6 reports the estimates and Section 7 relates them to observable characteristics. Section 8 reports the counterfactual analysis, and Section 9 concludes.

2 Background and Literature

We position the paper relative to four bodies of literature. The first is personalized pricing and selection markets, in which the terms a consumer faces depend on what she and others reveal, so participation is itself an equilibrium object. The second is privacy economics, in which the dominant empirical model treats disclosure as a one-equation choice over a scalar latent privacy preference. The third is bounded rationality and the cognitive hierarchy model, which provides the behavioral apparatus for heterogeneity in depth of reasoning but has been applied almost exclusively to firm-side decisions in well-defined games. The fourth is welfare in behavioral economics and adverse selection, which provides the framework for evaluating

welfare when consumers' perceived and realized payoffs differ, and which is essential for interpreting the redistributive content of information policy.

A. Personalized Pricing, Endogenous Participation, and Selection

Firms use the data consumers disclose to set prices. In the classical models, a firm that learns a consumer's type charges the individual accordingly, so disclosure transfers surplus from the consumer to the firm (Taylor, 2004; Conitzer et al., 2012), and consumers respond by managing what they reveal (Bonatti and Cisternas, 2020; Montes et al., 2019; Dubé and Misra, 2023). When disclosure is voluntary, participation is itself an equilibrium object: a seller draws inferences from silence, so favorable types disclose, the residual pool deteriorates, and disclosure unravels (Ali et al., 2023). Giving consumers control over what they reveal can therefore lower their welfare (Ichihashi, 2020). Each consumer's decision to reveal changes the inference applied to everyone who does not.

This makes data sharing a selection market in the sense of Einav and Finkelstein (2011): the terms a participant faces depend on the average type of those who participate (Akerlof, 1970; Einav et al., 2010; Veiga and Weyl, 2016). The literature has established that consumer sophistication is not unambiguously good. Gabaix and Laibson (2006) show that firms shroud add-on prices because naive consumers cross-subsidize sophisticated ones, and educating them removes the subsidy rather than creating surplus. Brown et al. (2010) examine the mechanism in field experiments on shrouded shipping charges. The implication is that informing consumers need not help them. Grubb and Osborne (2015) estimate that bill-shock alerts, which correct consumers' overconfidence about their own usage, would have lowered average consumer welfare since carriers raise prices in response. Handel (2013) finds that removing inertia in health-plan choice worsens adverse selection enough to reduce welfare, and Handel et al. (2019) show that correcting information frictions can amplify selection. The evidence is not universal: Agarwal et al. (2015) find that the CARD Act's fee restrictions lowered borrowing costs with no offsetting rise in interest charges. Whether debiasing

consumers helps them is an empirical question about how the market responds.

We contribute to this literature in three ways. First, in the personalized-pricing models, every consumer is rational, so the extent of unraveling is governed by preference heterogeneity alone. We let consumers differ in how deeply they reason and estimate that distribution, which turns the extent of unraveling into an estimable function of rationality rather than an assumption. Second, in insurance and other selection markets, the terms on offer are posted, and a sophisticated consumer need only read them. In data-sharing markets, the information posted is limited: how intensively an organization acts on what it learns is not provided, so the sophisticated consumer has to form a belief about who else shares. That belief is what we recover. Third, the welfare effect of sophistication operates through redistribution among members. Naive consumers do not merely cross-subsidize sophisticated ones; their sharing is what keeps the pool worth joining. Whether average welfare rises depends on the distribution of cognitive levels, which the model recovers, and on the welfare relevance of personalization, which it does not.

B. Economics of Privacy and Data Markets

The classical economics of privacy decomposes consumers' willingness to share data into two motives. The first is an intrinsic taste for secrecy that is invariant to use, privacy as a "final good" in the sense of Hermalin and Katz (2006). The second is an instrumental concern about the consequences of disclosure, the concealment motive of Stigler (1961) and Posner (1981), formalized as privacy as an "intermediate good" by Hermalin and Katz (2006) and decomposed empirically by Lin (2022). The empirical economics of privacy has shifted from a willingness-to-pay tradition (Acquisti et al., 2016) toward structural decomposition of the privacy preference, with a focus on the *privacy paradox*: the systematic gap between consumers' stated concerns and their revealed sharing behavior (Athey et al., 2017). Field data (Tucker, 2014; Goldfarb and Tucker, 2011) showed that consumer responses are highly context-dependent across data-collector identities.

Our paper builds on the strand that decomposes the privacy preference into structural components. Lin (2022) separates an intrinsic taste for secrecy from an instrumental concern about how a firm uses the data. The theoretical anchor for the instrumental margin is Bergemann et al. (2022), whose model of social-data markets shows how downstream firms extract surplus from consumer information and motivates the heterogeneous strategic use across organizations that we estimate. Choi et al. (2019) show that one consumer's disclosure can reveal information about others, and Kummer and Schulte (2019) that consumers perceive firms' use of data as varying across contexts. We extend this body of work by introducing a component to the instrumental margin, a strategic-use cost that depends on the equilibrium composition of the contributor pool, and identify it jointly with the cognitive-level distribution from a within-person trust panel.

Modern privacy economics also recognizes externalities as a distinct category of welfare-relevant primitive. Goldfarb and Que (2023) provide a discussion on the data externalities, in which one consumer's data sharing reveals information about non-consenting third parties through firm inference. Que (2026) indicates an externality in the sequential data-consenting environment, where an app's data-collecting strategy impacts the consent rates of downstream apps. We identify a selection externality in voluntary data-sharing. Because an organization's use of data depends on the average type of those who share with it, each consumer's decision to share shifts the pool that every other sharer faces, an adverse-selection force in the sense of Akerlof (1970). Unlike the inference externality borne by non-consenters (Choi et al., 2019; Goldfarb and Que, 2023), this externality falls on the consenters themselves, and its magnitude is bounded by the population's cognitive distribution rather than fixed at the full-rationality limit.

The closest privacy-specific behavioral work is Benndorf et al. (2015), who use level-k to explain incomplete unraveling in a laboratory disclosure game. We extend that insight by bringing bounded reasoning to a structurally estimated model of disclosure and by tracing its market and policy consequences. Estimating heterogeneity in strategic sophistication

allows us to quantify how reasoning shapes selection into disclosure, the composition of the disclosing pool, and thus the equilibrium extent of unraveling. We then use the estimated model to study the welfare effects of policies that inform consumers about the consequences of disclosure. Because such policies change not only individual disclosure decisions but also who discloses, their welfare effects operate through endogenous pool composition as well as direct individual choice.

C. Bounded Rationality and CH Model

The cognitive-hierarchy (CH) and level-k frameworks model heterogeneity in strategic reasoning by organizing players into a hierarchy parameterized by a single cognitive intensity (Stahl and Wilson, 1995; Nagel, 1995; Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Crawford and Iriberri, 2007; Crawford et al., 2013). Three features make the framework well suited to our setting. A single parameter orders players by sophistication. The truncated-Poisson assumption delivers a unique behavioral prediction even with multiple Nash equilibria, which is essential for welfare counterfactuals. And the implied posterior of higher types over lower ones can be computed forward without fixed-point iteration, which makes structural likelihood feasible at our sample size.

Empirical applications of CH have proliferated in firm-side and product-market settings. Goldfarb and Xiao (2011) estimate CH on entry games in US local telephone markets and link estimated sophistication to managerial characteristics. Brown et al. (2012, 2013) use it to explain why film studios cold-open low-quality movies. Hortaçsu et al. (2019) embed CH in structural bidding for the Texas electricity market and use the cognitive distribution to evaluate efficiency consequences of mergers. This is the closest welfare analog to ours, since it links the distribution of strategic ability to a market-level welfare object. Gill and Prowse (2016) show with experimental panel data that cognitive ability predicts equilibrium-closer choices, and warn that external proxies for sophistication underperform structural recovery. We avoid this concern by recovering sophistication from the behavior the model is designed

to explain. Batzilis et al. (2019) document substantial non-equilibrium reasoning on roughly a million online rock-paper-scissors observations.

The broader behavioral-IO literature (Heidhues and Kőszegi, 2018; Gabaix, 2019) embeds heterogeneity in attention, inference, and reasoning into consumer-demand models, and behavioral public finance (Mullainathan et al., 2012) characterizes welfare consequences of policies that operate through choice. We know of no other paper that applies CH to consumer privacy disclosure. The natural strategic primitive in our setting, inference about the average data value of others who share, is absent from the manager-side and product-market applications that dominate the existing field literature.

D. Welfare in Behavioral Economics and Adverse Selection

Counterfactual evaluation in a model of bounded rationality requires a welfare framework that does not equate consumer choice with consumer welfare. Bernheim and Rangel (2009) provide the canonical such framework: when perceived and realized payoffs differ, because the consumer fails to anticipate a cost or overweights a benefit, the welfare-relevant object is the realized payoff. The framework has been applied to insurance plan choice (Handel and Kolstad, 2015), unemployment insurance (Spinnewijn, 2015), and corrective taxation (Allcott et al., 2019), in each case yielding welfare conclusions that diverge from the standard inclusive-value metric of Williams (1977); Small and Rosen (1981). We apply it to consumer privacy. A naive consumer who shares without internalizing the strategic-use cost still pays it, so information improves her decisions given the pool; the re-sorting of pools that follows can nonetheless leave her worse off.

The equilibrium channel of our welfare decomposition is the structural analog of the Akerlof (1970) adverse-selection mechanism in a market with bounded rationality. In Akerlof's classical environment, full rationality drives the market to a corner: high-quality sellers withdraw when the price reflects average quality, the pool deteriorates, and trade can collapse. The same logic operates in our voluntary data-sharing market. Classical signaling under full

rationality (Spence, 1973) can support a separating equilibrium, and the lemons literature lets the unraveling proceed to a corner. In our setting the unraveling is bounded by the cognitive-hierarchy structure and by residual heterogeneity in private privacy preferences, which anchors some sharing even at the extractive end of the rank. The market unravels in the Akerlof direction but not all the way, and the size of the unraveling is an explicit function of the rationality assumption that the classical literature treats as fixed.

Our methodological contribution is to embed the Akerlof unraveling inside a cognitive-hierarchy structure that makes its magnitude an estimable function of rationality, and to evaluate the resulting welfare in the Bernheim–Rangel framework. The two embeddings produce a decomposition of the welfare effects into a decision channel that operates only on consumers who were making mistakes and an equilibrium channel that operates on every consumer who continues to share. The decision channel is a gain that accrues to the initially naive consumers whose choices the policy corrects; the equilibrium channel lowers the payoffs of continuing sharers as pools re-sort. Which channel dominates depends on how much of the dependence of payoffs on type is real surplus rather than a transfer among members: under pure redistribution the equilibrium cost exceeds the decision gain, and the sign reverses only when roughly half of that dependence is welfare relevant (Section 8.9). The welfare-relevant statistics for a consent-based regime are therefore the joint distribution of preferences and cognitive levels in the population and the welfare relevance of personalization.

3 Empirical Setting

This section describes the LISS data-donation study (Section 3.1), and documents three facts that motivate our model of strategic disclosure. First, mean trust in the nine types of data-acquiring organizations varies sharply. This variation is summarized by an institutional *harm rank*, along which trust declines linearly within person (Section 3.2). The second fact shows that, on average, respondents who better understand what sharing entails are more sensitive

to which organization would receive their data: their trust declines more steeply along the harm rank (Section 3.3). The third shows that the direction of this response depends on where a respondent's own data place her relative to others: richer respondents' trust falls faster along the harm rank and poorer respondents' trust more slowly, and this pattern is concentrated among the tertiary-educated (Section 3.4). Together they motivate the model in Section 4, in which consumers differ in how far they reason about the consequences of sharing.

The reduced-form evidence in this section documents three main facts from the survey. Section 7 connects the cognitive levels estimated by the model to a respondent's comprehension score and her other demographic characteristics. Section 3.5 closes by quantifying how stated trust relates to actual sharing, which we observe at one type of organization.

3.1 Data and Sample Construction

The data come from a 2024 module fielded in the LISS Panel (CentERdata, 2024). The LISS Panel is a probability-based online panel of households in the Netherlands, sampled from the population register and designed to be representative of the Dutch adult population.

Trust in organization types. Respondents first completed a questionnaire on their data-sharing attitudes. It measures their trust in nine types of organizations with respect to the handling of personal data. The question read:

> A lot of personal information is collected nowadays. To what extent do you trust that the following organizations will not share your personal information with others?

- Companies that create apps for smartphones
- Online shops, such as Amazon or bol.com
- Social media companies, such as Facebook, Twitter and LinkedIn
- Technology companies, such as Google
- Public pollsters and market research agencies, such as Ipsos or Kantar

- Scientific researchers
- Government bodies, such as the municipal or national government
- Official statistics institutes, such as Statistics Netherlands (CBS)
- Your internet provider

Responses were recorded on a four-point scale (from 1, "do not trust at all", to 4, "fully trust"). The question was asked in a randomized order across respondents. The four-point scale is as follows:

1. I do not trust them at all
2. I barely trust them
3. I trust them a bit
4. I fully trust them

Measuring comprehension. The survey includes a seven-item *comprehension quiz*, asked before the consent request, that tests whether the respondent understood the study and what sharing would entail: which data would be requested, how they would be filtered, whether names would be removed, whether she could review the package before donating, and whether she could be identified from it. The score each person gets is the number of correct answers she had among seven questions, ranging from 0 to 7. Among the respondents who answered the quiz, the mean is 3.72. Throughout this section we use the score as a descriptive measure of how well a respondent understands what sharing entails. It is not the model's measure of strategic sophistication. The structural estimation in Section 4 recovers each respondent's cognitive level from her trust profile and her sharing decision alone; the model does not use the quiz score. Section 7 examines how the recovered cognitive levels relate to the quiz score and to the participants' demographics.

Analysis sample. At the end of the questionnaire, respondents were asked to donate their WhatsApp Data Download Packages (DDPs) to the research team. These packages consisted of a group-chat-history package and an account-information package. The study

was originally designed for Corten et al. (2025). A total of 4,800 panel members were invited and 3,598 responded, corresponding to a response rate of 75.0%. We construct the main analysis sample by combining respondents' stated trust in each organization type, with their revealed sharing decision of their WhatsApp data. We restrict the sample to respondents who (a) own a smartphone; (b) use WhatsApp; (c) use their phone for messaging; and (d) reached the consent question. The resulting analysis sample contains $N = 3{,}132$ respondents. Trust ratings, income, education, and the sharing decision are observed for all of them; Table 1 reports item non-response on the other variables.

Key outcomes. We observe the following outcomes for each respondent. The stated outcome is a trust rating $T_{is} \in \{1, 2, 3, 4\}$ for the nine types of data-acquiring organizations. A higher rating indicates greater trust in that type of organization. The nine types span the range of data collectors a consumer would encounter: scientific researchers (academia), Statistics Netherlands (CBS), government entities, internet providers, market-research firms, online retailers, mobile-app companies, technology companies such as Google, and social-media platforms. The revealed outcome is a binary, consequential sharing decision $d_i \in \{0, 1\}$ of their WhatsApp message metadata at the host research organization, the researcher team. The respondent decided whether to consent to the research team on her data. In the analysis sample, 26.1% of respondents consent.[3] The binary decision we use is consent to share data, that captures whether the respondent agreed to donate her WhatsApp data when asked. Completing the donation required further technical steps for example exporting and uploading the data packages that we do not want to contaminate our measure.

[3]In incentivized experiments in which participants decide whether to release personal data to the research team, 29–50% share absent compensation (Lin, 2022). Participation in probability-sample data-collection pipelines spans a similar range: 13–17% for the Understanding Society expenditure app (Jäckle et al., 2023) and 46% for GPS-sensor sharing in a Dutch general-population survey (Struminskaya et al., 2021).

Table 1. Summary Statistics of the Sample

	Mean	SD	Min	Median	Max	*N*
Panel A. Outcomes and cognitive inputs						
Consent to donate data	0.26	0.44	0	0	1	3,132
Comprehension score (0–7)	3.72	2.42	0	4	7	2,741
Smartphone proficiency (1=beginner … 5=advanced)	3.28	0.99	1	3	5	2,851
Digital literacy index (1–5)	4.14	0.98	1	4.4	5	3,116
Panel B. Contextual trust, by institution (1–4)						
Scientific researchers	3.04	0.74	1	3	4	3,132
Statistics bureau (CBS)	3.06	0.75	1	3	4	3,132
Government	2.95	0.77	1	3	4	3,132
Internet providers	2.73	0.77	1	3	4	3,132
Market research firms	2.38	0.83	1	2	4	3,132
Online shops	2.45	0.80	1	3	4	3,132
App companies	2.27	0.75	1	2	4	3,132
Tech companies	2.20	0.82	1	2	4	3,132
Social media	1.97	0.80	1	2	4	3,132
Panel C. WhatsApp usage (self-reported)						
WhatsApp usage frequency (1=daily … 7=never)	1.48	0.97	1	1	7	2,810
Number of WhatsApp contacts	93	152	0	50	1,000	2,787
Number of WhatsApp groups	8.61	11.47	0	5	80	2,787
Panel D. Data-sharing outcomes and characteristics						
Shared chat history (=1 if yes)	0.14	0.34	0	0	1	3,132
Shared WhatsApp account info (=1 if yes)	0.10	0.31	0	0	1	3,132
# WhatsApp groups in shared file (of donors)	52	53	1	36	339	280
# WhatsApp contacts in shared file (of donors)	275	229	6	211	1,783	280

Notes: Analysis sample (N = 3,132): respondents who own a smartphone, use WhatsApp, use it for messaging, and reached the consent question. Per-row N counts non-missing observations. The 391 respondents who skipped the comprehension quiz are retained since the structural model does not use the quiz results. Self-reported contacts and groups are top-coded at 1,000 and 80. Zero is coded as value zero, not missing.

Key variables on individuals. Beyond the stated and revealed outcomes, we use the following key variables to anchor the reduced-form evidence and the structural model. The first is a set of demographic variables: monthly income, age, education, household structure and size, urbanization, and occupation. The second is the digital-footprint of an individual, which approximates the data value of each individual: smartphone proficiency, digital-literacy, and WhatsApp usage intensity, which is measured by frequency of use, number of contacts, and number of groups. Table 1 reports summary statistics for the analysis sample.

3.2 Motivating Fact 1: Trust Varies Sharply Across Organizations

Respondents distinguish the organizations that might acquire their data. They largely agree on the rating ranking. Figure 1 plots the mean trust rating $\bar{T}_s$ for the nine organization types in the sample. $\bar{T}_s$ falls from 3.06 at Statistics Netherlands to 1.97 at social-media platforms, which shows a wide spread. It spans 1.09 points, roughly 1.7 times the average within-respondent standard deviation. This cross-organization variation, and the trust ranking, is the dominant feature in the data. The rest of the section builds on this pattern.

We summarize this ranking with an institutional *harm rank*. This harm rank runs from 1 for the most-trusted type (Statistics Netherlands) to 9 for the least-trusted (social-media platforms), in the same order as in Figure 1. The harm rank is a summary of how trust varies across organizations. Its order matches the usual regulatory and economic view on which types of organizations use personal data more intensively. It is not driven by how we built the sample.[4]

Figure 1 compares averages across respondents. Since every respondent rates all nine types, we are able to measure the decline in trust within-person. We stack each respondent's nine ratings T_{is} into a person-by-organization panel and estimate the within-person slope of

[4]Table 13 in Appendix A reports the means for all 3,598 respondents who reached the trust module, and the ordering is essentially unchanged.

trust on the harm rank:

$$y_{is} \;=\; \alpha_i \;+\; \beta \cdot \text{HarmRank}_s \;+\; \varepsilon_{is}, \tag{1}$$

where $y_{is} \in \{1, 2, 3, 4\}$ is respondent i's *trust* in organization s. HarmRank_s is the harm rank of that organization type, ranging from 1 to 9. α_i is the individual fixed effect. The fixed effect absorbs everything about the respondent that does not vary across organizations, including the general privacy preference. Therefore, β is identified from within-person variation alone. Standard errors are clustered at the respondent level. The estimate is $\hat{\beta} = -0.143$, which means a one-step increase in the harm rank is associated with a trust lower by 0.143 points on average. The linear specification is a close first-order approximation. Replacing the linear term with eight rank dummies gives an almost perfectly monotonic decline (Appendix B).

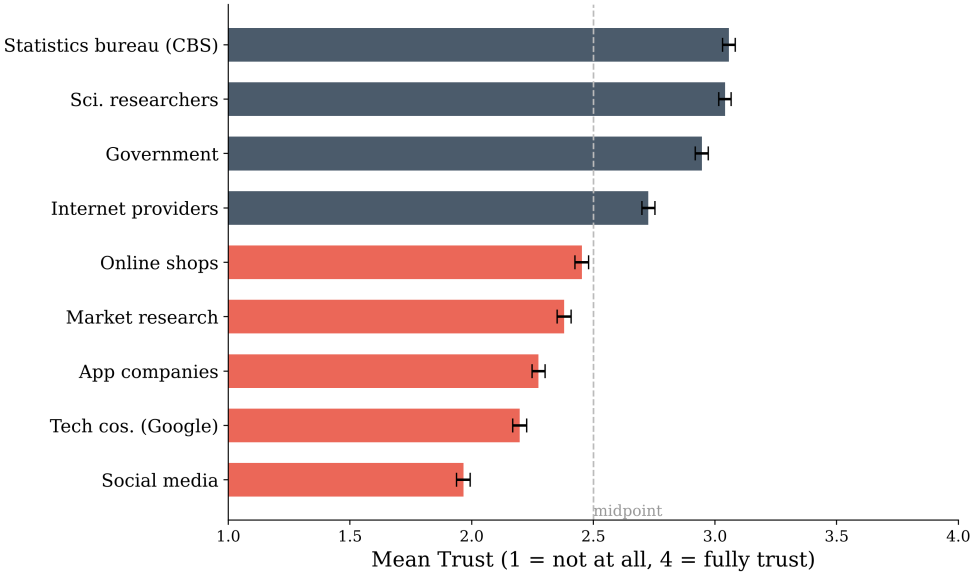


Figure 1. Sample-mean trust rating $\bar{T}_s$ at each of the nine organizations on the four-point Likert scale (1 = "not at all" to 4 = "fully trust"). Bars report sample means. Error bars are 95% confidence intervals.

Figure 1 and equation (1) in this section both describe the consistent average rating ranks. Respondents differ in how steeply their trust falls along the harm rank. We turn to

these differences next.

3.3 Motivating Fact 2: Individuals Who Understand Sharing Perceive Institutional Harm More Sharply

The gap widens with the respondents who best understood what sharing entails. Panel A of Figure 2 splits the mean trust rating $\bar{T}_s$ by comprehension group (low 0–2, middle 3–5, high 6–7). The three groups rank the organizations in the same order, but the spread within each group differs. The gap between the most- and least-trusted types rises from 0.93 points in the low group to 1.09 in the middle group and 1.36 in the high group. Every pairwise difference is statistically significant.[5] High-comprehension respondents trust low-harm types slightly more, and high-harm types substantially less.

[5]For each group we take each respondent's own trust difference between that group's most- and least-trusted types and compare mean differences across groups (Welch tests; $n = 869$, 1,052, and 820). The widening is +0.16 scale points from low to middle (SE 0.045), +0.26 from middle to high (SE 0.044), and +0.43 from low to high (SE 0.047), all $p < 0.001$; equal gaps are jointly rejected ($F = 41.6$, $p < 0.001$).

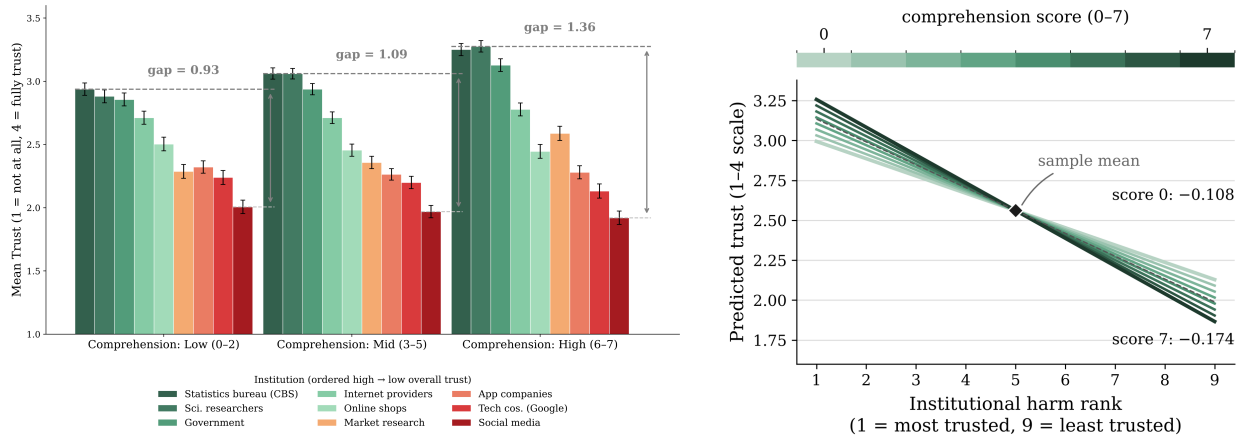


Figure 2. The relationship between trust and institutional harm, by comprehension score group. *Panel A*: sample-mean trust $\bar{T}_s$ by organization and comprehension group (low 0–2, middle 3–5, high 6–7), organizations ordered by overall mean trust. Error bars are 95% confidence intervals. *Panel B*: each line is the profile implied by equation (2) with comprehension score as the sole moderator (column (2) of Table 2), $\bar{y} + [\hat{\beta}_0 + \hat{\beta}_c\,(c - \bar{c})]\,(\text{HarmRank}_s - 5)$, at each score $c = 0, \ldots, 7$ (light to dark). Every line in this figure is anchored at the sample midpoint (rank 5, mean trust 2.56), so only the slopes, not the levels, are informative. The dashed line is the pooled slope of −0.143 from equation (1).

Panel A of Figure 2 shows the difference in comprehension levels in three coarse bins, but equation (1) imposes one slope on every respondent. We therefore let that slope vary with person-level characteristics:

$$y_{is} \;=\; \alpha_i \;+\; \beta_0 \cdot \text{HarmRank}_s \;+\; \sum_m \beta_m \cdot \big(\text{HarmRank}_s \times z_{i,m}\big) \;+\; \varepsilon_{is}, \tag{2}$$

where $\{z_{i,m}\}$ are proxies for understanding and skill from the survey responses, such as the comprehension score, the digital-literacy index, and an indicator for tertiary education. A negative β_m means that respondents scoring higher on $z_{i,m}$ trust higher-harm types less, relative to their own baseline. Table 2 reports the baseline of equation (1) (column (1)), each moderator alone (columns (2)–(4)), and all three jointly (column (5)).

The gradient is steeper for respondents who score higher on each proxy: the interaction terms of the proxies and harm rank are negative and statistically significant when entered

Table 2. Within-Person Trust Slope on Institutional Harm: Baseline and Heterogeneity by Comprehension and Skills

	(1) Baseline	(2) Het: Comprehension	(3) Het: Digital literacy	(4) Het: High education	(5) Het: All three
Dependent variable: Trust in organization (y_{is}, 1–4)					
Harm rank (centered)	-0.143***	-0.143***	-0.143***	-0.143***	-0.143***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Harm rank × Comprehension	—	-0.009***	—	—	-0.007***
		(0.001)			(0.001)
Harm rank × Digital literacy	—	—	-0.011***	—	-0.002
			(0.002)		(0.002)
Harm rank × High education	—	—	—	-0.068***	-0.062***
				(0.004)	(0.004)
Individual fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	28,044	28,044	28,044	28,044	28,044
R^2 (within)	0.295	0.302	0.297	0.311	0.315

Notes: Harm rank is the centered 1–9 ordering of organizations by sample-mean trust (rank 1 = lowest harm). Sophistication proxies are centered at their means. Digital literacy is on the mean-of-items (1–5) scale. A negative interaction means a steeper within-person decline in trust as harm rises. Column (2) is the specification behind panel B of Figure 2. The estimation sample drops the 16 respondents with missing digital literacy (3,116 respondents); the comprehension moderator is mean-imputed for the 378 respondents with a missing quiz score. Standard errors are clustered at the person level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

one at a time (columns (2)–(4)). When the proxy for understanding is comprehension scores, the estimate of the interaction term is −0.009: each additional quiz point is associated with a within-person slope steeper by 0.009, for each higher rank. Panel B of Figure 2 traces the implied tilt. The slope drops from −0.108 to −0.174 when the comprehension score changes from 0 to 7, roughly 60 percent steeper. Digital literacy and tertiary education enter negatively as well (columns (3)–(4)). This negative result does not rely on a linear specification in harm rank.[6]

Respondents who understand what sharing entails distinguish more sharply between organizations. A steeper slope means that a respondent's trust depends more on the organizational type, not her trust level in general, since the individual fixed effect absorbs a respondent's overall level of trust and privacy preference. These respondents may also simply know more about how commercial platforms use personal data. The slope in Table 2 is,

[6]Table 14 in Appendix B re-estimates the joint specification with harm discretized into high/low indicators, into three institutional clusters, and into rank dummies. The comprehension and education interactions are statistically significant in every functional form. The digital-literacy interaction is not.

however, an average over respondents with different data. In the model of Section 4, what an organization's use of the data means for a consumer depends on how her type compares with that of the other sharers, so the direction of her response depends on her position. We turn to this next.

3.4 Motivating Fact 3: The Response to Data Use Depends on the Value of One's Own Data

In the model, the consequence of sharing depends on how the consumer's type compares with the average type of the pool of consumers who share. We do not observe the pool in the reduced-form evidence, but we can measure where each respondent stands in the population. We do so in two ways. We first ask whether a respondent's distance from the population average, on several characteristics, steepens her trust gradient, and for whom. Distance measures how much a respondent's data reveal, but not in which direction. We then ask whether the direction of the response follows the sign of her income relative to the average, which is the comparison the model's payoff depends on.

We measure distance on several observed personal characteristics: for each continuous characteristic c, the distance is $\Delta_{i,c} = |x_{i,c} - \bar{x}_c|/s_{x_c}$, the standardized absolute deviation from the sample mean. For categorical characteristics, such as occupation, we use the rarity of the respondent's category. A respondent with $\Delta_{i,c} = 0$ reveals little by sharing their data, and one with $\Delta_{i,c} = 2$ reveals a lot. We test characteristics on *demographic distance* (i.e., age, log monthly net income, education, household size, number of children, urbanization, occupation rarity) and *digital-footprint distance* (i.e., WhatsApp contacts, groups, frequency, smartphone-activity count, digital literacy).

We extend equation (2) allowing two additional interaction terms: a) how being distant from the sample average changes the trust for each additional harm rank; and b) how

comprehension tilts the distance interactions:

$$y_{is} = \alpha_i + \beta_0 \,\mathrm{HarmRank}_s + \beta_c \left(\mathrm{HarmRank}_s \times \mathrm{Compr}_i\right) + \sum_m \beta_m \left(\mathrm{HarmRank}_s \times \Delta_{i,m}\right) + \sum_m \gamma_m \left(\mathrm{HarmRank}_s \times \Delta_{i,m} \times \mathrm{Compr}_i\right) + \varepsilon_{is}, \tag{3}$$

where Compr_i is the comprehension score. All characteristics (demographic and digital-footprint) are centered at their sample means. The individual fixed effect absorbs $\Delta_{i,m}$, Compr_i, and their product. Thus, the harm-rank interactions are identified. A negative β_m means that more distinctive respondents trust higher-harm organizations less, with comprehension level fixed. A negative γ_m means that this interaction strengthens with comprehension. Table 3 reports two-way estimates ($\gamma_m = \beta_c = 0$) in columns (1), (3), and (5), for the demographic dimensions, the digital-footprint dimensions, and all twelve, and the corresponding three-way specifications in columns (2), (4), and (6).[7]

Income and education distance steepen the within-person gradient, which shows negative and significant effects in every two-way column. A negative coefficient means that for a respondent who is far from the average, her trust falls faster as the harm rank rises. These are the closest analogs of the type in the model, which Section 4 measures by income. The digital-footprint dimensions are smaller and mixed in sign. Occupation rarity, the number of WhatsApp contacts, and the number of smartphone activities enter with positive coefficients, which flatten the gradient. We do not interpret these dimensions further.

The three-way columns show that the role of distance is affected by comprehension. The Harm × Compr. two-way interaction term shows that her slope becomes steeper with each additional quiz point, as in Motivating Fact 2. The distance interactions now compare a more distinctive respondent with the baseline respondent at their average score. The three-

[7]These are respondents with complete data on all twelve distance dimensions and a quiz score.

Table 3. Sophistication and the Value of One's Own Data: Within-Person Evidence

	(1)	(2)	(3)	(4)	(5)	(6)
	Demographic distance		Digital-footprint distance		All twelve dimensions	
	2-way	+3-way	2-way	+3-way	2-way	+3-way
Dependent variable: Trust in organization (y_{is}, 1–4)						
Harm	-0.1458***	-0.1456***	-0.1458***	-0.1439***	-0.1458***	-0.1440***
	(0.0023)	(0.0023)	(0.0023)	(0.0023)	(0.0023)	(0.0023)
Harm × Compr.	—	-0.0090***	—	-0.0088***	—	-0.0084***
		(0.0009)		(0.0010)		(0.0009)
Panel A. Distance × Harm						
Demographic distance						
$\Delta_{\text{age}}\times$ Harm	-0.0059	-0.0064	—	—	-0.0074	-0.0071
	(0.0044)	(0.0043)			(0.0045)	(0.0045)
$\Delta_{\text{income}}\times$ Harm	-0.0143***	-0.0100***	—	—	-0.0130***	-0.0091**
	(0.0040)	(0.0039)			(0.0041)	(0.0039)
$\Delta_{\text{educ}}\times$ Harm	-0.0091**	-0.0114***	—	—	-0.0102**	-0.0122***
	(0.0040)	(0.0039)			(0.0040)	(0.0039)
$\Delta_{\text{hh size}}\times$ Harm	-0.0023	-0.0008	—	—	-0.0019	-0.0006
	(0.0063)	(0.0062)			(0.0063)	(0.0062)
$\Delta_{\text{children}}\times$ Harm	0.0074	0.0098	—	—	0.0083	0.0098
	(0.0070)	(0.0070)			(0.0070)	(0.0069)
$\Delta_{\text{urban}}\times$ Harm	-0.0080*	-0.0076*	—	—	-0.0077*	-0.0077*
	(0.0046)	(0.0045)			(0.0046)	(0.0045)
$\Delta_{\text{occup}}\times$ Harm	0.0992***	0.0883***	—	—	0.0934***	0.0842***
	(0.0154)	(0.0152)			(0.0154)	(0.0152)
Digital-footprint distance						
$\Delta_{\text{contacts}}\times$ Harm	—	—	0.0098**	0.0073*	0.0102***	0.0078**
			(0.0038)	(0.0038)	(0.0038)	(0.0038)
$\Delta_{\text{groups}}\times$ Harm	—	—	-0.0085***	-0.0043	-0.0064**	-0.0030
			(0.0028)	(0.0029)	(0.0028)	(0.0031)
$\Delta_{\text{frequency}}\times$ Harm	—	—	0.0048	0.0063	0.0044	0.0067
			(0.0043)	(0.0048)	(0.0044)	(0.0050)
$\Delta_{\text{activities}}\times$ Harm	—	—	0.0117***	0.0122***	0.0113***	0.0119***
			(0.0042)	(0.0043)	(0.0042)	(0.0043)
$\Delta_{\text{dig. lit.}}\times$ Harm	—	—	0.0006	0.0030	0.0010	0.0039
			(0.0045)	(0.0045)	(0.0044)	(0.0045)
Panel B. Distance × Harm × Comprehension						
Demographic distance						
$\Delta_{\text{age}}\times$ Harm × Compr.	—	0.0039**	—	—	—	0.0035*
		(0.0018)				(0.0019)
$\Delta_{\text{income}}\times$ Harm × Compr.	—	-0.0031**	—	—	—	-0.0028*
		(0.0016)				(0.0016)
$\Delta_{\text{educ}}\times$ Harm × Compr.	—	-0.0045***	—	—	—	-0.0047***
		(0.0016)				(0.0016)
$\Delta_{\text{hh size}}\times$ Harm × Compr.	—	-0.0014	—	—	—	-0.0010
		(0.0026)				(0.0026)
$\Delta_{\text{children}}\times$ Harm × Compr.	—	-0.0007	—	—	—	-0.0006
		(0.0029)				(0.0029)
$\Delta_{\text{urban}}\times$ Harm × Compr.	—	-0.0006	—	—	—	-0.0004
		(0.0018)				(0.0018)
$\Delta_{\text{occup}}\times$ Harm × Compr.	—	-0.0008	—	—	—	-0.0033
		(0.0063)				(0.0063)
Digital-footprint distance						
$\Delta_{\text{contacts}}\times$ Harm × Compr.	—	—	—	-0.0023	—	-0.0021
				(0.0015)		(0.0015)
$\Delta_{\text{groups}}\times$ Harm × Compr.	—	—	—	-0.0006	—	-0.0006
				(0.0012)		(0.0012)
$\Delta_{\text{frequency}}\times$ Harm × Compr.	—	—	—	0.0041**	—	0.0037*
				(0.0019)		(0.0020)
$\Delta_{\text{activities}}\times$ Harm × Compr.	—	—	—	0.0068***	—	0.0064***
				(0.0017)		(0.0017)
$\Delta_{\text{dig. lit.}}\times$ Harm × Compr.	—	—	—	0.0011	—	0.0014
				(0.0017)		(0.0016)
Individual fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	24,561	24,561	24,561	24,561	24,561	24,561
R^2 (within)	0.309	0.317	0.307	0.315	0.311	0.320

Notes: Standard errors clustered at individual level. Harm is the centered 1–9 ordering of organizations by sample-mean trust (rank 1 = Statistics Netherlands). $\Delta_c = |x_{i,c} - \bar{x}_c|/s_{x_c}$ is respondent i's standardized distance from the sample mean on characteristic c (category rarity for categorical variables such as occupation); Compr. is the 0–7 quiz score. Contacts enter as $\log(1+\text{contacts})$; groups are top-coded at 80; activities count the 15 smartphone uses in the survey. The sample consists of the 2,729 respondents with all twelve characteristics and the quiz score. In Panel A, a coefficient is the change in the trust–harm slope per unit of Δ_c (i.e. distance from the mean data value) at mean comprehension. In Panel B, a coefficient is the change in that per quiz point. A negative coefficient means trust falls faster with harm. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

way terms show how this comparison changes as the comprehension rises: a negative term means that distance steepens the slope more for respondents with higher scores. The triple interactions are jointly significant everywhere.[8] For income and education, both two-way and three-way interaction terms are negative. The extra steepness that comes with distance grows with the quiz score. This means, at the bottom of the quiz scale, a respondent far from the average income or education has about the same slope as a respondent of average income or education. Among the individuals on the higher-comprehension end, the same distance is associated with a steeper slope.[9]

Distinctive respondents thus respond more strongly to the organization's use of data, especially when they understand what sharing entails. The model makes a sharper, signed prediction. A consumer who reasons about the pool compares her type with the average type in the pool. At an organization that charges more on higher types, a consumer above the average expects to give up more than the average, and one below the average expects to give up less. A naive consumer does not make this comparison, so comprehensive proxies, such as income, shift her trust by the same amount at every organization. Under the model, in this setting, income changes trust differently across organizations only for consumers who reason about the pool. Since commercial organizations rank high in harm, this reasoning should make trust fall faster along the harm rank for richer consumers and more slowly for poorer ones. Table 4 tests this prediction. It adds to equation (2) the interaction of the harm rank with standardized log income, the type in the model, and the interactions of this term with tertiary education and with the comprehension score.

The signed pattern is present in the data. In column (1), an increase in income steepens the within-person decline in trust. Columns (2) and (3) show that the pattern is stronger

[8] Cluster-robust Wald $p = 0.009$ in column (2), $p < 0.001$ in columns (4) and (6).

[9] This is not constrained to the linear specification. Replacing the continuous score in column (2) with the three groups in Figure 2, the estimates of $\Delta_{\text{income}} \times$Harm is 0.000, -0.012, and -0.019 across the low, middle, and high groups. The estimates of $\Delta_{\text{educ}} \times$Harm is $+0.006$, -0.016, and -0.023. The high-group estimates differ significantly from the low-group estimates ($p = 0.041$ for income and $p = 0.004$ for education).

Table 4. Position Relative to Others: The Within-Person Trust Slope on Harm by Income

	(1)	(2)	(3)	(4)	(5)
Dependent variable: Trust in organization (y_{is}, 1–4)					
Harm rank (centered)	-0.143***	-0.143***	-0.141***	-0.144***	-0.141***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Harm rank × Income	-0.023***	-0.012***	-0.013***	-0.022***	-0.014***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Harm rank × High education	—	-0.059***	-0.057***	—	-0.051***
		(0.005)	(0.005)		(0.005)
Harm rank × Income × High education	—	—	-0.013**	—	-0.014***
			(0.005)		(0.005)
Harm rank × Comprehension	—	—	—	-0.009***	-0.007***
				(0.001)	(0.001)
Harm rank × Income × Comprehension	—	—	—	0.003***	0.003***
				(0.001)	(0.001)
Implied Harm rank × Income, from column (3)					
Without tertiary education			-0.008**		
			(0.003)		
With tertiary education			-0.021***		
			(0.004)		
Individual fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	28,044	28,044	28,044	28,044	28,044
R^2 (within)	0.302	0.313	0.313	0.309	0.318

Notes: Within-person OLS of trust on the centered harm rank and its interactions. Income is standardized log net monthly income, the type g_i of the structural model, centered at the sample mean. A negative Harm rank × Income coefficient means that trust falls faster along the harm rank for richer respondents and more slowly for poorer respondents. High education is an indicator for tertiary education; comprehension is the 0–7 quiz score. All moderators are centered at their sample means, so lower-order coefficients are evaluated at the mean. The sample is that of Table 2: the 3,116 respondents with non-missing digital literacy, with the comprehension score mean-imputed for the 378 respondents with a missing quiz score. The implied slopes are linear combinations of the column-(3) coefficients. Standard errors are clustered at the person level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

among the tertiary-educated ones. Columns (4) and (5) show that the triple interaction with the quiz score is small, positive, and significant. The comprehension score is thus associated with a stronger average response to the organization (Fact 2), but not with a sharper income-dependent direction of the response, while education is associated with both. This anticipates the structural estimates, in which the cognitive level rises with income and education but not with the quiz score once these are accounted for. The harm rank is a summary of the organizations.

3.5 Quantifying the Gap: Stated Trust and Actual Sharing

Our analysis so far focuses on stated trust in the survey. However, the structural model concerns actual disclosure decisions. We observe such a decision for one organization type: academic researchers. We estimate how stated trust translates into actual data sharing, and we use the estimates as the bridge from the surveyed trust panel to the structural disclosure model. Let $T_{is} \in \{1, 2, 3, 4\}$ denote person i's self-reported trust in organization s, mapped to the unit interval by $\text{Trust}_{is} = (T_{is} - 1)/3 \in [0, 1]$. In the context of sharing data to academic researchers, $s =$ researchers, we estimate the linear-probability model

$$Y_{\text{share},i} \;=\; \alpha \;+\; \beta \cdot \text{Trust}_i \;+\; \boldsymbol{\gamma}'\mathbf{X}_i \;+\; \varepsilon_i, \tag{4}$$

where Trust_i is normalized trust in researchers and $\mathbf{X}_i$ denotes demographic characteristics and survey measures of comprehension and skill. The coefficient β measures the trust pass-through, which is the change in actual sharing probability associated with an increase in normalized trust from 0 to 1, holding other characteristics fixed. The estimates in Table 15 of Appendix C show that stated trust strongly predicts actual sharing. A 0-to-1 increase in normalized trust is associated with a 23-percentage-point higher probability of sharing after controlling for demographics. This association decreases to 13 percentage points when we additionally control for survey measures of comprehension and skill. Appendix C reports the

full specification and summarizes the residual variation in sharing across respondents that stated trust does not explain.

4 A Cognitive-Hierarchy Model of Data Disclosure

This section presents the model of heterogeneous strategic reasoning that we take to the data. Individuals decide whether to share their personal data with each of several types of organization. The consequences of sharing depend on the pool of individuals whose data reach the organization, and individuals differ in how far they reason about the choices of others that form this pool. Organizations run data programs whose terms depend on the composition of the pool of participants. We model the decisions as a simultaneous game: in most data markets, individuals know little about others' sharing decisions when they make their own.

4.1 Environment

There are N individuals, indexed by $i = 1, \ldots, N$, and $S = 9$ organization types, indexed by $s = 1, \ldots, S$. For each organization type, individual i decides whether to share her data, $d_{is} \in \{0, 1\}$. An individual is characterized by a type g_i and a privacy cost θ_i. The type is what her data reveal: it is her private information before she shares, and it is observed by the organization if she does.[10] The privacy cost θ_i is not observed by the econometrician. It may vary with type and is common across organizations; organization-specific consequences of sharing enter through the organization's data use, described below. We specify

$$\theta_i = \beta_g \, g_i + u_i, \qquad u_i \sim \mathcal{N}(0, \sigma_\theta^2). \tag{5}$$

[10] In our baseline estimation, g_i is standardized log net monthly income. Income is the characteristic most directly tied to willingness to pay, and hence to personalized pricing and targeting.

4.2 Organizations

Each organization runs a data program, and an individual joins the program by sharing her data. An individual who does not share receives the organization's standard terms, which do not depend on her type or on her decision not to share. We normalize her payoff from the standard terms to zero.

The organization may use the data it collects to improve its service or to set its prices. We assume that, without an individual's data, the organization does not condition her treatment on her type. Aggregate effects of others' sharing, such as better services, improved algorithms, or changes in general prices, are part of the standard terms available to everyone and are absorbed in the normalization of the outside option. This is the standard assumption in selection markets, where the payoff of the uninsured does not depend on who buys insurance (Einav et al., 2010). It is natural when personalization requires identifiable individual-level data and non-participants receive a posted product, a default service, or a uniform public treatment. The assumption requires only that the organization commits, before individuals decide, to terms for non-participants that do not vary with the subsequent composition of the non-sharing pool. The organization may anticipate selection into participation, but it does not revise an individual's terms on the basis of her refusal. This commitment describes posted-price, default-service, and regulated-service environments, in which terms are not recalculated from the identities of those who decline to share.

This assumption distinguishes our setting from the disclosure models of Grossman (1981), Milgrom (1981), and Ali et al. (2023), in which the organization treats an individual who withholds her data as the average individual who withholds. In those models, other individuals matter through the pool of non-sharers, and this is what drives unraveling. In our model, non-sharers receive a standard option because of technological, contractual, legal, or organizational constraints on personalization. The disclosure decision is the same, but other individuals matter through the pool of sharers: a sharer's payoff depends on who else shares. This is adverse selection among participants, as in Akerlof (1970). In the real disclosure deci-

sion we observe, sharing data with researchers, refusal ends the interaction, so the restriction holds directly. Non-disclosure preserves the status quo, and the organization-specific value of sharing relative to the status quo is captured by v_s.

This structure describes many opt-in data programs. A customer who does not join a loyalty program can still buy at posted prices, while a member reveals her purchase history, becomes eligible for personalized offers and program rewards, and is exposed to personalized treatment. A user who declines tracking by a website can still use the standard version of the service, while a consenting user is exposed to data-dependent personalization and targeting. A person who does not contribute data to a research or public-sector program continues to receive the standard service, whereas contributors generate benefits shared among participants and are exposed to data-enabled benefits or harm.

Data use. When individual i shares, the organization observes g_i and uses it, for example to set personalized prices or to target offers. The resulting transfer from this sharer individual to the organization is

$$t_s(g_i) = r_s + \psi_s \, g_i, \tag{6}$$

where r_s is the transfer from a sharer of average type and ψ_s measures how the transfer varies with the sharer's type. When $\psi_s > 0$, the organization takes more from higher types, as with pricing or targeting based on willingness to pay (Taylor, 2004; Acquisti and Varian, 2005). When $\psi_s < 0$, it takes more from lower types, for example by offering better service to high types. If the organization also returns part of the value of the data through personalized benefits that vary with type, ψ_s is the slope of the transfer net of the slope of these benefits. The sign of ψ_s therefore tells us which types lose from personalization, and its magnitude tells us by how much.

Common benefit. Every sharer also receives a common benefit B_s, such as rewards, free or improved service, or the public services that the data help provide. In addition to the

type-specific transfer, each member's data generate a value v_s for the organization that does not come at the member's expense. The organization sets B_s so that the program breaks even,

$$\sum_{j \in \mathcal{D}_s} \big[t_s(g_j) + v_s - B_s \big] = 0 \quad \Longrightarrow \quad B_s = r_s + v_s + \psi_s \, \bar{g}_s^*, \tag{7}$$

where $\mathcal{D}_s = \{j : d_{js} = 1\}$ is the set of data sharers and $\bar{g}_s^*$ is their average type. Break-even is the outcome of competition for users (Armstrong, 2006) or, for public organizations, of budget balance. Equation (7) is the average-cost pricing condition of selection markets (Akerlof, 1970; Einav et al., 2010), applied to the common benefit: the terms that every member receives are set by the average member.

A loyalty program illustrates the mechanism. Members share their purchase data, which the firm uses for personalized offers and prices, and every member earns the same reward rate, financed by what the program earns. Members whose data are most valuable to the firm give up the most through personalized pricing. If they leave, the program earns less, the reward rate falls, and every remaining member is worse off. An individual deciding whether to join must therefore anticipate who else will join the pool.

Payoff from sharing. A member receives B_s, pays $t_s(g_i)$, and bears her privacy cost θ_i. Her payoff from sharing, relative to not sharing, is

$$\eta_{is}^* = B_s - t_s(g_i) - \theta_i = v_s - \theta_i - \psi_s\big(g_i - \bar{g}_s^*\big). \tag{8}$$

The average transfer r_s cancels because the program returns to its members, on average, what it takes from them. What remains is the difference between what individual i gives up and what the average member gives up. When $\psi_s > 0$, a member with $g_i > \bar{g}_s^*$ subsidizes the other members, so high types have weaker incentives to share, and when they leave, B_s falls and the remaining members are worse off. When $\psi_s = 0$, the payoff from sharing does not depend on who else shares.

Two features of the model are worth emphasizing. First, individual i does not care directly about other individuals' decisions or utilities. Other individuals matter only because their decisions determine the pool, and the pool determines her payoff. Second, the organization commits to its program, summarized by v_s, ψ_s, and the break-even rule, before individuals decide, and applies it to the realized pool. It makes no decision after the pool is realized, so we do not assign it a level of reasoning. In the counterfactual simulations, we hold v_s and ψ_s fixed and let B_s adjust with the composition of the pool.[11]

We measure payoffs in the units of the trust ratings described in Section 4.4. With i.i.d. type-I extreme-value shocks, an individual who decides on payoff η_{is} shares with organization s with probability

$$\Pr(d_{is} = 1 \mid \eta_{is}) = \Lambda\big(\gamma_0 + \gamma_1\, \eta_{is}\big). \tag{9}$$

Here Λ is the logistic function. An individual who knew the pool would decide on $\eta_{is} = \eta^*_{is}$; Section 4.3 specifies the payoff that individuals perceive at each cognitive level. The parameter γ_1 converts payoffs from the scale of trust ratings into the scale of actual sharing decisions, and γ_0 captures costs or benefits of the act of sharing that are common to all individuals and not reflected in trust.

4.3 Cognitive Levels and Beliefs

Individuals do not observe $\bar{g}^*_s$ when they decide. They forecast it, and they differ in how far they reason about the choices of others that form it. Following Camerer et al. (2004), each individual has a cognitive level $k \in \{0, 1, \ldots, K\}$, drawn from a truncated Poisson distribution with parameter τ,

$$\pi_k = \frac{\tau^k / k!}{\sum_{\ell=0}^{K} \tau^\ell / \ell!}, \tag{10}$$

with $K = 2$.

[11] In estimation, $\bar{g}^*_s$ is computed over the other members. The exact payoff is $v_s - \theta_i - \psi_s \frac{n_s - 1}{n_s}(g_i - \bar{g}^*_{-i,s})$, where n_s is the number of members; in our data $(n_s - 1)/n_s > 0.997$.

Level 0. A level-0 individual ignores that her payoff depends on how her type compares with the pool and behaves as if she were an average type member. She knows that organizations differ, through v_s, but she does not reason about her own position in the pool that determines the personalization consequences. Her perceived payoff is

$$\eta_{is}^{(0)} = v_s - \theta_i. \tag{11}$$

Level $k \geq 1$. A level-k individual believes that other individuals have levels $h = 0, \ldots, k-1$, in proportion to their population frequencies,

$$\omega_{h|k} = \frac{\pi_h}{\sum_{\ell=0}^{k-1} \pi_\ell}, \qquad h = 0, \ldots, k-1. \tag{12}$$

She does not observe other individuals' privacy costs, so she forecasts that a level-h individual j shares with probability

$$\bar{P}_{js}^{(h)} = \int \Lambda\big(\gamma_0 + \gamma_1 \, \eta_{js}^{(h)}(u)\big) \, \phi(u) \, du, \tag{13}$$

the sharing rule (9) evaluated at j's perceived payoff, where $\eta_{js}^{(h)}(u)$ is that payoff when $u_j = u$. Her forecast of the average type of the pool is the weighted average

$$\bar{g}_{is}^{(k)} = \frac{\sum_{j \neq i} g_j \, W_{js}^{(k)}}{\sum_{j \neq i} W_{js}^{(k)}}, \qquad W_{js}^{(k)} = \sum_{h=0}^{k-1} \omega_{h|k} \, \bar{P}_{js}^{(h)}, \tag{14}$$

and her perceived payoff is

$$\eta_{is}^{(k)} = v_s - \theta_i - \psi_s\big(g_i - \bar{g}_{is}^{(k)}\big). \tag{15}$$

A level-1 individual thus expects the pool to be formed by level-0 individuals, and a level-2 individual expects a mixture of level-0 and level-1 individuals. We use a cognitive hierarchy rather than a level-k model because the object that individuals forecast is a pool

formed by a heterogeneous population. In a level-k model, a level-2 individual behaves as if all other individuals are level 1; in a cognitive hierarchy, she expects a mixture of levels 0 and 1, as in equation (14). These beliefs imply different forecasts of the pool and therefore different perceived costs of revealing one's type. Because consumers forecast with the same sharing rule that generates the pool, a consumer's forecast is wrong only because of what she believes about the cognitive levels of others. The cognitive hierarchy also summarizes the distribution of reasoning with the single parameter τ, which allows us to study an increase in sophistication as a change in τ.

4.4 Observation Equations and Likelihood

We observe two outcomes. The first is a four-point trust rating $T_{is} \in \{1, 2, 3, 4\}$ for each of the nine organizations. The second is an actual data-sharing decision $d_i \in \{0, 1\}$, that captures whether the respondent consented to donate her WhatsApp data to the researchers, whom we denote by $s = R$. We treat both as measures of the same perceived payoff. The trust question asks whether they trust the organization will keep the respondent's information from others. We treat the rating as a monotone index of the perceived consequences of sharing with the organization, and read ψ_s as the personalization portion of these consequences. This is reasonable because within-person differences remove each respondent's use of the trust scale and her general taste for privacy.

The trust rating of individual i for organization s follows an ordered logit,

$$\Pr(T_{is} = m \mid \eta_{is}^{(k)}) = \Lambda\big(c_m - \eta_{is}^{(k)}\big) - \Lambda\big(c_{m-1} - \eta_{is}^{(k)}\big), \qquad m = 1, \ldots, 4, \tag{16}$$

with $c_0 = -\infty$, $c_4 = +\infty$, and $c_1 < c_2 < c_3$. The sharing decision follows the sharing rule (9) at the researchers, evaluated at the perceived payoff,

$$\Pr(d_i = 1 \mid \eta_{iR}^{(k)}) = \Lambda\big(\gamma_0 + \gamma_1\, \eta_{iR}^{(k)}\big). \tag{17}$$

This is the same rule that individuals use to forecast the sharing decisions of others. Because we observe actual sharing at only one organization, we model that the mapping from stated trust to actual sharing through parameters (γ_0, γ_1), which is common across organizations. Once the organization-specific consequences of disclosure are summarized by η_{is}, the gap between stated trust and actual sharing does not depend on the organization and only depends on the common mapping. Under this assumption, the donation decision identifies the scale of actual sharing, and differences across organizations are identified from within-person variation in trust.

Individual i's contribution to the likelihood mixes over cognitive levels and integrates over their privacy cost,

$$\mathcal{L}_i = \int \sum_{k=0}^{K} \pi_k \left[\prod_{s=1}^{S} \Pr\big(T_{is} \mid \eta_{is}^{(k)}(u)\big) \right] \Pr\big(d_i \mid \eta_{iR}^{(k)}(u)\big)\, \phi(u)\, du. \tag{18}$$

We approximate the integral by Gauss–Hermite quadrature with 21 nodes; the log-likelihood changes by less than one point with 31 or 41 nodes.

4.5 Welfare

The model separates the payoff on which an individual decides from the payoff she receives. A level-k individual decides on her perceived payoff $\eta_{is}^{(k)}$, but if she shares she receives the realized payoff η_{is}^{*}, which depends on the actual pool. A level-0 individual, for example, shares without accounting for $\psi_s(g_i - \bar{g}_s^{*})$, but she bears it once she shares. We therefore evaluate welfare with the realized payoff, following the distinction between decision utility and experienced utility (Kahneman et al., 1997) and the behavioral welfare approach of Bernheim and Rangel (2009).

We measure welfare on the scale of actual sharing decisions. Let $V_{is}^{(k)} = \gamma_0 + \gamma_1 \eta_{is}^{(k)}$ be the perceived value of sharing, $q_{is}^{(k)} = \Lambda(V_{is}^{(k)})$ the resulting probability of sharing, and $V_{is}^{*} = \gamma_0 + \gamma_1 \eta_{is}^{*}$ the realized value. An individual who chooses according to $V_{is}^{(k)}$ but receives

V^*_{is} has expected utility

$$u^{(k)}_{is} = q^{(k)}_{is}\, V^*_{is} + H\big(q^{(k)}_{is}\big), \qquad H(q) = -q\log q - (1-q)\log(1-q), \tag{19}$$

where $H(q)$ is the expected value of the logit taste shocks (Leggett, 2002; Allcott, 2013; Train, 2015); we interpret the logit disturbances as idiosyncratic utility shocks that enter both decision and experienced utility. When beliefs are correct, $V^{(k)}_{is} = V^*_{is}$ and equation (19) reduces to the standard logit surplus $\log(1 + e^{V^*_{is}})$ (Small and Rosen, 1981). Otherwise, $u^{(k)}_{is}$ falls short of this amount by a nonnegative decision loss, which is zero only when $V^{(k)}_{is} = V^*_{is}$. Welfare is measured in units of the log-odds of sharing.

The realized pool is formed by actual decisions. The probability that individual j shares with s is $p_{js} = \sum_k \pi_k \int \Lambda\big(V^{(k)}_{js}(u)\big)\phi(u)\,du$, and the realized average type of the other members is

$$\bar{g}^*_{-i,s} = \frac{\sum_{j\neq i} p_{js}\, g_j}{\sum_{j\neq i} p_{js}}, \tag{20}$$

which enters the realized payoff η^*_{is} through equation (8). The welfare of individual i is $U_i = \sum_s \sum_k \pi_k \int u^{(k)}_{is}(u)\,\phi(u)\,du$. In the counterfactuals of Section 8, we replace the population weights by each respondent's posterior distribution over her cognitive level and privacy cost, given her trust ratings and sharing decision.

Implications for welfare. Because the program breaks even, the personalized terms and the common benefit cancel when summed over members: $\sum_{j\in\mathcal{D}_s} \psi_s(g_j - \bar{g}^*_s) = 0$. Personalization is therefore a transfer among members, and total welfare depends on who shares only through $v_s - \theta_j$ and the taste shocks. A level-0 individual decides on exactly this social value, so the allocation in which every individual is naive is efficient. Sophistication adds a private incentive to sort on type, which moves participation away from this allocation: the efficiency cost of strategic disclosure is that participation no longer tracks $v_s - \theta_i$. Information that lets individuals sort has private value but no social value (Hirshleifer, 1971). An increase

in sophistication improves each individual's decision given the pool, but it also changes the pool, and in our estimates the second effect dominates. This conclusion rests on treating the type gradient ψ_s as redistribution; Section 8.9 reports welfare when part of it reflects real surplus.

The trust ratings and the sharing decision identify how individuals perceive the consequences of disclosure, summarized by v_s and ψ_s. They do not identify the organizations' costs, profits, or actual conduct. Our welfare analysis therefore concerns individuals.

5 Identification and Estimation

The parameters to be estimated are the organization values v_s, the strategic-use coefficients ψ_s, the privacy-cost parameters (β_g, σ_θ), the cognitive intensity τ, the trust cutpoints (c_1, c_2, c_3), and the sharing scale (γ_0, γ_1). We normalize $v_s = 0$ at the statistics bureau. Identification comes from three sources of variation.

Variation in average trust across organizations within respondents. Every respondent rates all nine organizations. Differences in mean trust across organizations within respondents identify the differences $v_s - v_{s'}$, and the frequencies of the four trust categories identify the cutpoints. The privacy cost is common across organizations, so the correlation of a respondent's ratings across organizations, after netting out v_s and the income component of θ_i, identifies σ_θ.

Variation in type within organizations. Within an organization, the slope of the latent index in g_i is $-\beta_g$ for a level-0 respondent, who ignores the pool, and $-(\beta_g + \psi_s)$ for a respondent of level $k \geq 1$. The component of the slope that is common to all organizations identifies β_g. The component that differs across organizations identifies ψ_s together with the share of respondents who respond to the pool. A population of naive respondents would produce the same slope at every organization; the more the slope varies across organizations with the sign and size of ψ_s, the more mass the data place on levels $k \geq 1$. The split between

levels 1 and 2 is identified from the different pool forecasts of the two levels, $\bar{g}_s^{(1)}$ and $\bar{g}_s^{(2)}$, which shift the level of the index differently across organizations. The truncated-Poisson form ties the three level shares to the single parameter τ.

The sharing decision. The relation between the perceived payoff at the researchers and the donation decision identifies (γ_0, γ_1): γ_1 is the response of the log-odds of sharing to the payoff, and γ_0 the level. The same parameters enter individuals' forecasts of others' sharing, but these forecasts depend on them only through relative sharing probabilities, so the donation decision is their main source of identification.

Estimation. We estimate all parameters jointly by maximum likelihood, equation (18), on the panel of 3,132 respondents and 28,188 respondent–organization observations. The integral over u_i is computed by Gauss–Hermite quadrature with 21 nodes. We maximize the likelihood with L-BFGS-B from five random starting values and one additional starting value; all six starts converge to the same optimum, with log-likelihood $-28{,}224.02$.

We truncate the hierarchy at $K = 2$. Empirical applications of the cognitive hierarchy find that behavior is concentrated at shallow depths of reasoning (Camerer et al., 2004; Crawford et al., 2013), and our estimates place less than four percent of the population at level 2. The truncation also has little effect on the counterfactual path. Allowing for more levels does not change the results. With $K = 3$ or $K = 4$, less than half a percent of respondents are at levels above two.

Inference. Standard errors are from a respondent-clustered pairs bootstrap with $B = 50$ replications. Standard errors of transformed parameters, such as τ and the level shares π_k, are computed from the bootstrap distribution of the transformed parameters. Confidence intervals for the counterfactual statistics are calculated by recomputing the counterfactual at each bootstrap draw of the parameters.

6 Results

6.1 Cognitive Levels

Table 5 reports the estimates. The cognitive intensity is $\hat{\tau} = 0.322$ (standard error 0.023), which implies that 72.8 percent of respondents are naive (level 0), 23.5 percent reason one step ahead, and 3.8 percent reason two steps ahead. Most consumers decide whether to share without reasoning about who else shares.

Table 5. Structural estimates

	Revealed share (joint)	
	est.	(SE)
β_g	0.133	(0.059)
σ_θ	2.017	(0.047)
mean ψ_s	-0.192	(0.181)
range ψ_s (min, max)	(-3.40, +2.78)	
Cognitive intensity τ	0.322	(0.023)
share at level 0, π_0	0.728	(0.016)
share at level 1, π_1	0.235	(0.011)
share at level 2, π_2	0.038	(0.004)
θ_i s.d. across respondents	1.889	
θ_i IQR (p25, p75)	(-1.36, +1.15)	
cutpoint c_1	-5.167	(0.092)
cutpoint c_2	-2.355	(0.070)
cutpoint c_3	1.976	(0.075)
γ_0 (donate intercept)	-1.117	(0.051)
γ_1 (donate slope on η)	0.177	(0.026)
log-likelihood	-28224.0	
respondents N	3,132	
organization-respondent obs	28,188	

Notes. Maximum-likelihood estimates of the model of Section 4, with organization-specific v_s and ψ_s reported in Table 6. The privacy cost is integrated by Gauss–Hermite quadrature with 21 nodes. Standard errors in parentheses are from a respondent-clustered bootstrap with 50 replications; standard errors of τ, π_k and σ_θ are computed from the bootstrap distribution of these transformed parameters. θ_i summaries refer to posterior means.

The large mass at level 0 makes the cognitive-hierarchy belief structure consequential. A

level-2 consumer believes that 76 percent of others are naive and 24 percent reason one step ahead. In a level-k model she would instead believe that all others reason one step ahead. Because naive consumers ignore the composition of the pool and level-1 consumers respond to it, the two models imply different forecasts of the pool that sophisticated consumers face.

6.2 Strategic Use of Data

Table 6 reports the organization-level estimates. The values $\hat{v}_s$ decline from the statistics bureau (normalized to zero) to social media (-3.81), mirroring the ranking of mean trust in Section 3. The strategic-use coefficients do not follow this ranking. They are negative at the statistics bureau (-3.06), the researchers (-3.40), the government (-2.68) and market research firms (-2.04), and positive at the internet provider ($+0.54$), online shops ($+1.90$), app companies ($+1.48$), technology companies ($+2.78$) and social media ($+2.76$). Each coefficient is at least 2.7 standard errors from zero.

Table 6. Organization-level estimates

Organization	$\hat{v}_s$	SE	$\hat{\psi}_s$	SE	$\bar{g}_s^{(1)}$	$\bar{g}_s^{(2)}$
Statistics bureau	+0.00		-3.06	0.33	-0.017	+0.063
Researchers	-0.08	0.05	-3.40	0.30	-0.017	+0.071
Government	-0.45	0.04	-2.68	0.28	-0.018	+0.056
Internet provider	-1.19	0.06	+0.54	0.20	-0.018	-0.037
Online shops	-2.13	0.09	+1.90	0.20	-0.019	-0.092
Market research	-2.63	0.07	-2.04	0.23	-0.019	+0.043
App companies	-2.80	0.08	+1.48	0.19	-0.019	-0.078
Tech companies	-2.99	0.09	+2.78	0.24	-0.019	-0.135
Social media	-3.81	0.08	+2.76	0.25	-0.020	-0.140

Notes. Organizations are ordered by mean trust in the analysis sample. $\hat{v}_s$ is normalized to zero at the statistics bureau. $\bar{g}_s^{(1)}$ and $\bar{g}_s^{(2)}$ are the pool averages forecast by level-1 and level-2 consumers at the estimates. Pool averages are in standard deviations of log income, relative to the sample mean. Standard errors are from a respondent-clustered bootstrap with 50 replications.

The sign of ψ_s tells us which consumers expect to lose from being identified. At commercial organizations, consumers with higher income expect to give up more than the average

sharer, for example through pricing or targeting based on willingness to pay, so selection into these pools is adverse. At public and research organizations, they expect to gain relative to the average sharer, so selection is advantageous. Market research firms, sixth of nine in mean trust, belong to the second group: consumers trust them less than public institutions, but do not expect them to exploit what they learn about higher-income respondents. The organization's baseline appeal and its perceived use of individual data are thus separately identified, and they need not move together.

The magnitudes are large relative to the trust scale. At technology companies, a one-standard-deviation increase in income lowers the latent index of a sophisticated consumer by $\hat{\beta}_g + \hat{\psi}_s = 2.91$, compared with 0.13 for a naive consumer. This difference is comparable to the distance between adjacent trust categories ($\hat{c}_2 - \hat{c}_1 = 2.81$).

6.3 Privacy Costs and the Sharing Decision

The privacy cost rises with income, $\hat{\beta}_g = 0.133$ (standard error 0.059), but income explains less than one percent of its variance: the standard deviation of the residual is $\hat{\sigma}_\theta = 2.02$. Most heterogeneity in privacy preferences is unrelated to income.

The sharing scale is $\hat{\gamma}_1 = 0.177$ (standard error 0.026). A one-unit increase in the perceived payoff at the researchers raises the log-odds of donating by 0.177, or the probability of donating by about 3.4 percentage points at the sample donation rate of 26 percent. The intercept $\hat{\gamma}_0 = -1.117$ implies a donation probability of 25 percent at a perceived payoff of zero. Actual sharing therefore responds much less to perceived payoffs than stated trust does, which is why we measure welfare on the scale of actual decisions.

7 Post-Estimation Analysis: Who Are the Cognitive Types?

Section 6 recovers two latent objects for each respondent. First, a privacy preference, summarized by its posterior mean $\hat{\theta}_i = \mathbb{E}[\theta_i \mid \text{data}_i]$. Second, a cognitive level, represented by $(\hat{p}_{i,0}, \hat{p}_{i,1}, \hat{p}_{i,2})$ with $\hat{p}_{i,k} = \Pr(K_i = k \mid \text{data}_i)$, and summarized by $\bar{K}_i = \sum_k k\, \hat{p}_{i,k}$. Both are recovered from the same likelihood. Privacy preference is identified by the overall position of a respondent's trust profile, while the cognitive level is identified by how the income gradient of trust varies across organizations (Section 5). In this section, we explore how the two estimates relate to respondents' observed characteristics, using the demographic, attitudinal, and digital-experience variables in the LISS background data ($N = 3{,}132$).

The evidence in this section reveals a clear asymmetry: demographic characteristics predict a person's cognitive level, but not her privacy preference. $\bar{K}_i$ is significantly associated with income and tertiary education, and more weakly with age. Once these demographic factors are controlled, the comprehension quiz score adds little explanatory power. Naive individuals are thus more concentrated among respondents with fewer resources. In contrast, income and education explain almost none of the variation in $\hat{\theta}_i$. It is mostly correlated with self-reported privacy concern. This result provides external validation for the recovered privacy preference. The implication is that a planner can identify individuals who are more likely to be naive, but not who value privacy from demographic characteristics alone. Both patterns are consistent with findings in the cognitive-hierarchy literature (Goldfarb and Xiao, 2011; Gill and Prowse, 2016) and the privacy literature (Tucker, 2014; Lin, 2022).

7.1 One-Step Estimates: Cognitive Intensity and Characteristics

Regressions of posterior means on characteristics have two limitations. For many respondents, the trust ratings and the sharing decision carry little information about the cognitive level. Their posterior then stays close to the population distribution and the regressions

understate how much levels vary with characteristics. The regressions also treat the posteriors as known, although they are estimated. The usual way to handle both problems in latent-class models is to let characteristics shift the class probabilities inside the likelihood (Dayton and Macready, 1988).

We therefore estimate a version of the model in which the cognitive intensity varies with respondent characteristics,

$$\log \tau_i = \delta_0 + X_i'\delta, \tag{21}$$

and respondent i's level is drawn from the truncated Poisson distribution with parameter τ_i. Everything else is the same as in Section 4. A level-k respondent forms her beliefs about others from the population distribution of levels. X_i contains the characteristics used below, with continuous variables standardized and missing values set to the mean with a missing-value indicator. All parameters are estimated jointly by maximum likelihood, with standard errors computed from the outer product of the scores.

Table 7 reports the estimates. The characteristics help predict who is naive and who reasons further, with income and education as the strongest predictors. Richer and more educated respondents are less likely to be naive. Frequent WhatsApp users are also less likely to be naive. Women are less sophisticated. Once these characteristics are in the model, the procedural quiz score, digital literacy, smartphone skill, and stated privacy concern do not predict cognitive intensity. Adding the characteristics also leaves the population shares of the three levels close to those in the main model. The income gradient is steep: the average posterior probability of being naive is close to one in the bottom income decile and just under two-thirds in the top decile.

The rest of this section describes both latent objects in more detail using their posterior means.

Table 7. Cognitive intensity and respondent characteristics: one-step estimates

Covariate	$\hat{\delta}$	(s.e.)	$\Delta\pi_0$
Age	-0.081	(0.070)	+0.016
Education (CBS 1–6)	-0.091	(0.107)	+0.018
Log income	+0.400***	(0.085)	-0.092
Procedural quiz (0–7)	+0.053	(0.058)	-0.011
Digital literacy	-0.016	(0.069)	+0.003
Smartphone skill	-0.044	(0.067)	+0.009
Privacy concern	-0.059	(0.050)	+0.012
WhatsApp frequency (7 = daily)	+0.156**	(0.071)	-0.034
Household members	+0.277	(0.225)	-0.062
Children at home	-0.276	(0.198)	+0.053
Female	-0.283**	(0.116)	+0.054
Tertiary education	+0.545***	(0.202)	-0.129
Has partner	-0.396*	(0.206)	+0.073
Procedural quiz (0–7) missing	+0.217	(0.349)	-0.048
Digital literacy missing	-0.288	(0.761)	+0.055
Smartphone skill missing	-0.465	(0.477)	+0.084
WhatsApp frequency missing	+0.258	(0.562)	-0.057
Intercept (δ_0)	-1.118	(0.192)	
Mean (π_0, π_1, π_2)		(0.738, 0.219, 0.043)	
Log-likelihood (model without covariates)		-28156.4 (-28224.0)	
LR test $\delta = 0$		$\chi^2_{17} = 135.2$	

Notes. Maximum-likelihood estimates of equation (21) jointly with the model of Section 4. Continuous covariates are standardized. $\Delta\pi_0$ is the average change in the probability of being naive for a one-standard-deviation increase in a continuous covariate or a change from 0 to 1 in an indicator. Standard errors from the outer product of the scores. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

7.2 Correlates of the Cognitive Level $\bar{K}_i$

We use observed characteristics, measured skills, and stated preference to correlate with $\bar{K}_i$. We project both posterior objects on the same covariates. Income, education, a quadratic measure in age, and life-cycle position, are all variables researchers and planners could observe in administrative data. The comprehension score, digital literacy, and smartphone skill measure cognitive and technological skill. Privacy concern records a stated attitude, while WhatsApp frequency measures platform use.

We begin with the comprehension score, which measures how well respondents under-

stood the study and what sharing entailed. It is a strong moderator of the trust–harm gradient (Section 3.3). However, as Gill and Prowse (2016) point out, a survey measure of sophistication may reflect demographic differences rather than an independent reasoning ability.

Even before any controls, the raw data lean toward the demographic reading. Mean $\bar{K}_i$ increases with income mainly at the bottom of the income distribution, and with education only at the tertiary level (Appendix D.1). The estimated cognitive levels differ most among respondents with fewer resources.

Table 8. Correlates of the posterior cognitive level $\bar{K}_i$ (OLS, HC1-robust standard errors, regressors on their natural scale).

	(1) $\bar{K}_i$	(2) $\bar{K}_i$	(3) $\bar{K}_i$	(4) $\bar{K}_i$	(5) $\bar{K}_i$
Comprehension score (0–7)	+0.0044 (0.0029)	+0.0009 (0.0029)	–	−0.0006 (0.0030)	–
Privacy concern (1–7)	–	–	–	−0.0068 (0.0053)	−0.0065 (0.0052)
Digital literacy (rowmean 1–5)	–	–	–	−0.0035 (0.0089)	−0.0035 (0.0087)
Smartphone skill (1–5)	–	–	–	+0.0008 (0.0094)	+0.0008 (0.0093)
Log net income	–	+0.1115*** (0.0093)	+0.1112*** (0.0089)	+0.1099*** (0.0108)	+0.1073*** (0.0107)
Education (CBS 1–6)	–	−0.0050 (0.0079)	−0.0085 (0.0071)	−0.0088 (0.0083)	−0.0081 (0.0081)
Tertiary educ. (≥5)	–	+0.0501** (0.0219)	+0.0542*** (0.0204)	+0.0563** (0.0224)	+0.0586*** (0.0221)
Age (years)	–	–	–	+0.0039 (0.0024)	+0.0044* (0.0024)
Age^2	–	–	–	−0.0000* (0.0000)	−0.0000** (0.0000)
WhatsApp frequency (1 = never, 7 = daily)	–	–	–	+0.0125 (0.0080)	+0.0096 (0.0070)
Constant	+0.2951*** (0.0126)	−0.5343*** (0.0667)	−0.5151*** (0.0626)	−0.6234*** (0.0910)	−0.6055*** (0.0877)
N	2,741	2,741	3,132	2,738	2,806
R^2	0.0008	0.0492	0.0463	0.0525	0.0518

Notes: The dependent variable in every column is the posterior mean cognitive level $\bar{K}_i$. Column (1): comprehension score only; (2): adds income and education; (3): income and education without the comprehension score; (4): the full set of observables; (5): the full set without the comprehension score. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HC1-robust SE in parentheses. Each column is estimated on its own listwise-complete (maximal) sample; the comprehension score and smartphone skill carry the binding item non-response.

The association between the cognitive level and the comprehension score is weak, and it disappears once income and education are taken into account. Table 8 examines this relationship across five specifications. Column (1) includes only the comprehension score. Column (2) adds income and education, and column (3) drops comprehension while retaining these demographic variables. Columns (4) and (5) repeat the last two specifications with the full set of controls. On its own, the comprehension score has a small and insignificant coefficient (0.004); after controlling for income and education, it falls to 0.001. Meanwhile, both estimates of log income and tertiary education remain statistically significant across

specifications. Age has a weak inverted-U relationship with cognitive level, while the other skill, attitude, and usage measures add little explanatory power. Demographics are informative but have limited predictive power. The full model explains only about 5% of the variation in cognitive level.

Income and education primarily predict whether a respondent is above level 0, rather than whether they reach level 2. Appendix D.2 shows this by examining the two cumulative posterior probabilities, $\Pr(K_i \geq 1)$ and $\Pr(K_i \geq 2)$. The naive individuals are more concentrated at the bottom of the income distribution. This pattern is relevant for the welfare analysis in Section 8, where the counterfactual shifts mass out of level 0.

7.3 Correlates of the Preference Type $\hat{\theta}_i$

In the model, differences in privacy preferences explain why some consumers continue to share data even with organizations that make extensive use of it. We measure each respondent's privacy preference by the posterior mean $\hat{\theta}_i$. A higher $\hat{\theta}_i$ indicates greater reluctance to share across all organizations. We examine whether these estimated preferences align with respondents' stated privacy concerns and whether they vary systematically with observable characteristics.

Figure 3 shows substantial heterogeneity in estimated privacy preferences. The distribution is centered at mean -0.03 with a standard deviation of 1.89. There is substantial mass in both tails. Some respondents are highly willing to share data, while others are strongly protective of their privacy. Unlike cognitive level $\bar{K}_i$, mean $\hat{\theta}_i$ varies little across income or education groups, although it tends to decline with digital skill (Appendix D.1).

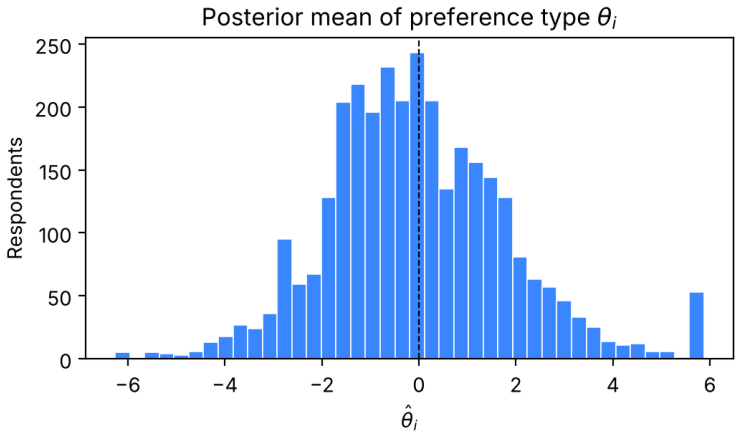


Figure 3. Marginal distribution of the posterior mean preference type $\hat{\theta}_i$, $N = 3{,}132$.

Stated privacy concern is the strongest observable correlate of the estimated privacy preference.[12] Since this question is not used in the structural estimation, the comparison provides an independent check of whether the recovered preference reflects respondents' attitudes toward privacy. Table 9 shows that a one-point increase in stated privacy concern is associated with a 0.31 increase in $\hat{\theta}_i$, even after controlling for income and education. Income and education explain almost none of the variation in estimated privacy preferences. Digital literacy and more frequent WhatsApp use are associated with greater willingness to share, but stated privacy concern remains the strongest predictor.

The strong relationship between stated privacy concern and estimated privacy preference offers one explanation for the privacy paradox (Athey et al., 2017). Privacy concern is strongly associated with $\hat{\theta}_i$, but not with cognitive level (Table 8). In the model, a consumer's decision to share with organization s depends on three factors: her underlying privacy preference θ_i, the consequences of sharing determined by the organization's data-use parameters v_s and ψ_s, and her cognitive level K_i, which determines whether she thinks about the consequences. Stated privacy concern measures only the first of these. However, disclosure behavior also depends on the organization's use of data and the respondent's cognitive level.

[12] In the survey, the question asks "How concerned are you about your privacy, in general?" on a 1–7 scale.

Two equally concerned respondents can therefore make different choices.

No observable group differs systematically in privacy preference. This is the second important claim of this section. Administrative data may help identify groups with lower cognitive levels, but income and education reveal little about who values privacy more. Respondents with high and low $\hat{\theta}_i$ are found across all demographic groups. This is consistent with modeling θ_i as a random effect. The individual variation is much larger than the part explained by observable characteristics (Sections 4 and 6). It also aligns with findings in the empirical privacy literature. Policies based on demographic characteristics are therefore unlikely to identify consumers with stronger privacy preferences.

Table 9. Correlates of the posterior preference type $\hat{\theta}_i$ (OLS, HC1-robust standard errors, regressors on their natural scale).

	(1) $\hat{\theta}_i$	(2) $\hat{\theta}_i$	(3) $\hat{\theta}_i$	(4) $\hat{\theta}_i$	(5) $\hat{\theta}_i$
Privacy concern (1–7)	+0.3117*** (0.0276)	+0.3114*** (0.0276)	–	+0.3277*** (0.0295)	–
Digital literacy (rowmean 1–5)	–	–	–	−0.2185*** (0.0473)	−0.2048*** (0.0495)
Comprehension score (0–7)	–	–	–	−0.0652*** (0.0153)	−0.0621*** (0.0157)
Smartphone skill (1–5)	–	–	–	−0.0089 (0.0482)	−0.0357 (0.0497)
Log net income	–	+0.0441 (0.0543)	+0.0635 (0.0563)	+0.0542 (0.0623)	+0.0509 (0.0646)
Education (CBS 1–6)	–	−0.0155 (0.0411)	−0.0045 (0.0424)	+0.0710 (0.0457)	+0.0807* (0.0472)
Tertiary educ. (≥5)	–	+0.0161 (0.1142)	−0.0256 (0.1172)	−0.0137 (0.1212)	−0.0385 (0.1249)
Age (years)	–	–	–	+0.0210 (0.0136)	+0.0347** (0.0138)
Age^2	–	–	–	−0.0003** (0.0001)	−0.0004*** (0.0001)
WhatsApp frequency (1 = never, 7 = daily)	–	–	–	−0.1068** (0.0459)	−0.1298*** (0.0457)
Constant	−1.3357*** (0.1221)	−1.6106*** (0.4012)	−0.4754 (0.4013)	−0.4299 (0.5318)	+0.7511 (0.5339)
N	3,132	3,132	3,132	2,738	2,738
R^2	0.0523	0.0525	0.0004	0.0800	0.0241

Notes: The dependent variable in every column is the posterior mean preference type $\hat{\theta}_i$. Column (1): privacy concern only; (2): adds income and education; (3): income and education without privacy concern; (4): the full set of observables; (5): the full set without privacy concern. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. HC1-robust SE in parentheses. Each column is estimated on its own listwise-complete (maximal) sample; the comprehension score and smartphone skill carry the binding item non-response.

7.4 Discussion

Demographic characteristics help predict how consumers reason about data use. They do not predict who minds it. Figure 4 compares these relationships on a standardized scale. Income, tertiary education, and, more weakly, age are associated with $\bar{K}_i$ but have little relationship with $\hat{\theta}_i$. In contrast, privacy concern and skill measures are associated with $\hat{\theta}_i$,

but have little relationship with $\bar{K}_i$.

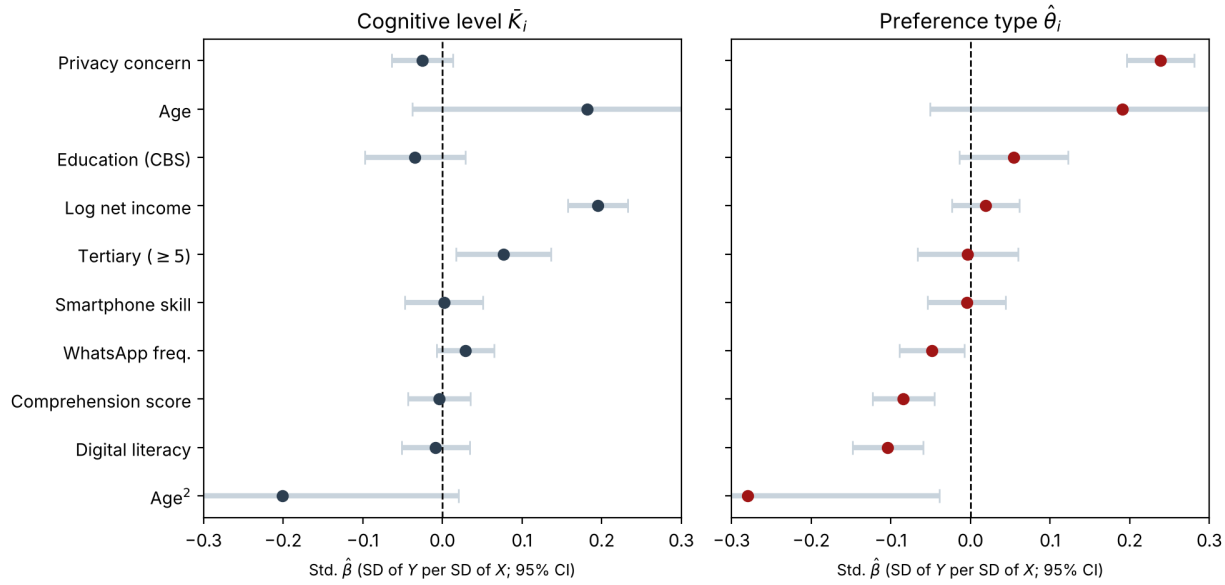


Figure 4. Standardized coefficients from regressions of the two posterior estimates on the same set of observable characteristics. Both outcomes and regressors are standardized, so each coefficient measures the change in the outcome (in standard deviations) associated with a one-standard-deviation increase in the regressor. *Left:* the cognitive level $\bar{K}_i$. *Right:* the preference type $\hat{\theta}_i$. Age and age^2 should be interpreted jointly.

This distinction is consistent with the model. Privacy preference is identified from the *location* of a respondent's trust profile, while the cognitive level is identified from how its income gradient varies across organizations (Section 5). These findings are consistent with previous research showing that cognitive sophistication is associated with observable characteristics (Goldfarb and Xiao, 2011; Gill and Prowse, 2016), while privacy preferences are difficult to predict from demographics (Tucker, 2014; Lin, 2022). Still, the correlations with self-reported measures provide supporting evidence rather than proof of the structure.

This distinction matters for policy. Demographic characteristics help identify groups with more level-0 consumers, but they reveal little about these consumers' privacy preferences. Section 8 examines the welfare implications of this information gap.

8 Counterfactual: Increasing Sophistication

We use the estimated model to simulate a policy that increases strategic sophistication, such as a disclosure or literacy intervention that leads consumers to reason about their personalization consequences. We first trace how the composition of each organization's sharer pool depends on the population's sophistication and compare it with the rational-expectations equilibrium (Section 8.1). We then evaluate the welfare effect of moving the population from the estimated $\hat{\tau}$ to $\tau_{\text{cf}} = 3$, a point on this path. At $\tau_{\text{cf}} = 3$, the shares at levels 0, 1 and 2 are 12, 35 and 53 percent.

The welfare analysis is conditional on the market structure of Section 4. We estimate the distribution of cognitive levels and the perceived consequences of sharing. In our baseline model, we assume that organizations run break-even programs whose common benefit depends on the composition of the pool, and that consumers' perceptions of each organization's use of data are correct. Later, we investigate how welfare can depend on the real surplus created by personalization.

8.1 Sophistication Path and Rational-Expectations Benchmark

In the classical lemons market, the extent of unraveling follows from the assumption that all participants are rational (Akerlof, 1970). In our model, the state variable of the unraveling process is the realized sharer-pool average $\bar{g}_s^*$ of equation (20). Only consumers of level $k \geq 1$ respond to the composition of the pool, and level-0 consumers decide as if they were average members. The cognitive intensity τ therefore determines what fraction of the population takes part in the unraveling, and the map $\tau \mapsto \bar{g}_s^*(\tau)$ traces it out. For each τ, we move each respondent's cognitive level from the estimated distribution $\pi(\hat{\tau})$ to $\pi(\tau)$ by a monotone coupling, so that no respondent becomes less sophisticated, hold all other parameters at their estimates, and solve for the realized pools.

The benchmark for this path is the rational-expectations equilibrium, in which beliefs

coincide with the realized pool: p_{is} and $\bar{g}_s^{\mathrm{RE}}$ jointly solve

$$p_{is} = \Lambda\big(\gamma_0 + \gamma_1 \left[\, v_s - \theta_i - \psi_s(g_i - \bar{g}_s^{\mathrm{RE}}) \,\right]\big), \qquad \bar{g}_s^{\mathrm{RE}} = \frac{\sum_j p_{js}\, g_j}{\sum_j p_{js}}. \tag{22}$$

The rational consumer is not a type in our hierarchy: as in the cognitive-hierarchy literature, equilibrium play arises only as the limit of the belief iteration (Camerer et al., 2004). Because we truncate the hierarchy at $K = 2$, the large-τ limit of the path need not coincide with the fixed point (22). In our estimates the difference is negligible: at $\tau = 30$ every pool average is within 0.006 of its rational-expectations value. Even under rational expectations the market does not unravel to a corner, because the logit taste shocks and the dispersion of θ_i sustain interior sharing at every organization.

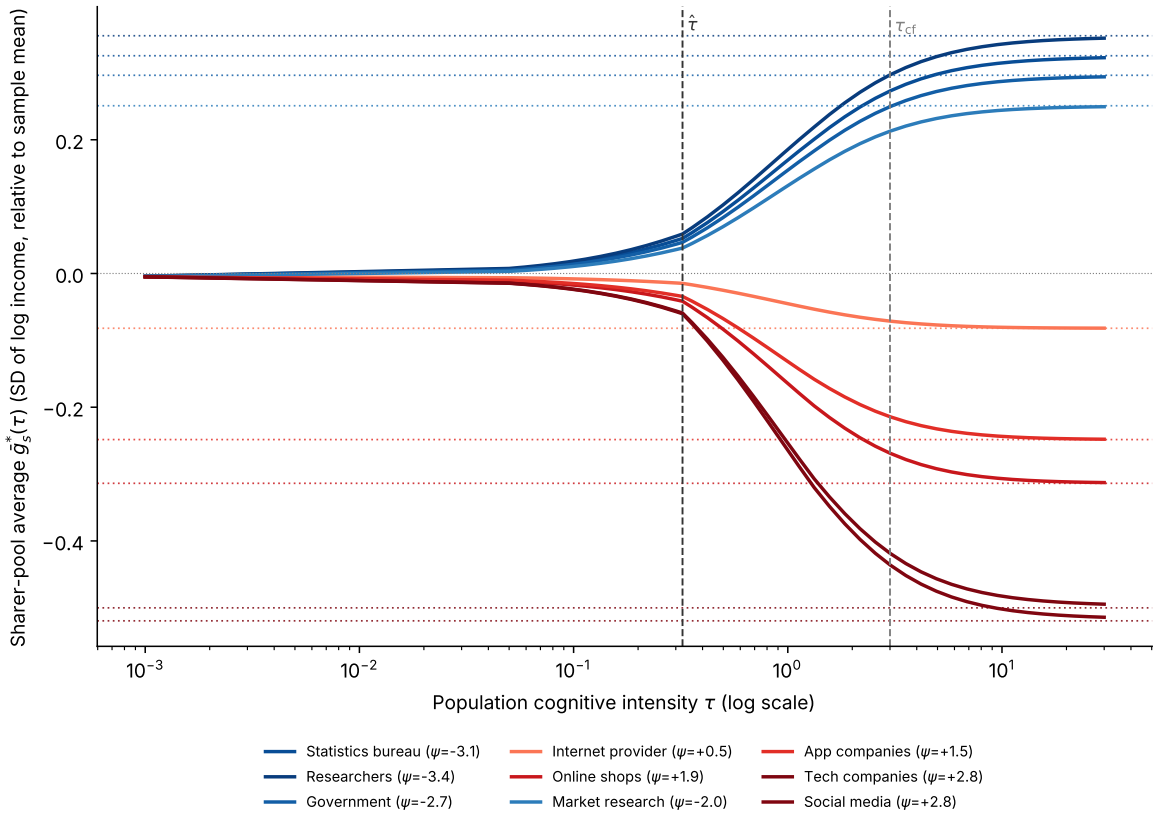


Figure 5. Sharer pools as a function of population sophistication. Each solid curve is the realized sharer-pool average $\bar{g}_s^*(\tau)$ at one organization, computed on a grid of the cognitive intensity τ with all other parameters at their estimates. Red curves are organizations with $\hat{\psi}_s > 0$ (adverse selection), blue curves organizations with $\hat{\psi}_s < 0$ (advantageous selection). Dotted horizontal lines are the rational-expectations benchmarks from equation (22). Vertical dashed lines mark the estimate $\hat{\tau} = 0.32$ and the counterfactual $\tau_{\text{cf}} = 3$.

Figure 5 reports the result. When $\tau \to 0$ the population is entirely naive, and every pool is close to the population average, because privacy costs are only weakly correlated with type. As τ rises, high-g consumers withdraw from the organizations with $\hat{\psi}_s > 0$, whose pools fall toward the rational-expectations benchmark, and they move toward the organizations with $\hat{\psi}_s < 0$, whose pools rise. The estimated population, at $\hat{\tau} = 0.32$, is close to the naive end of the path.

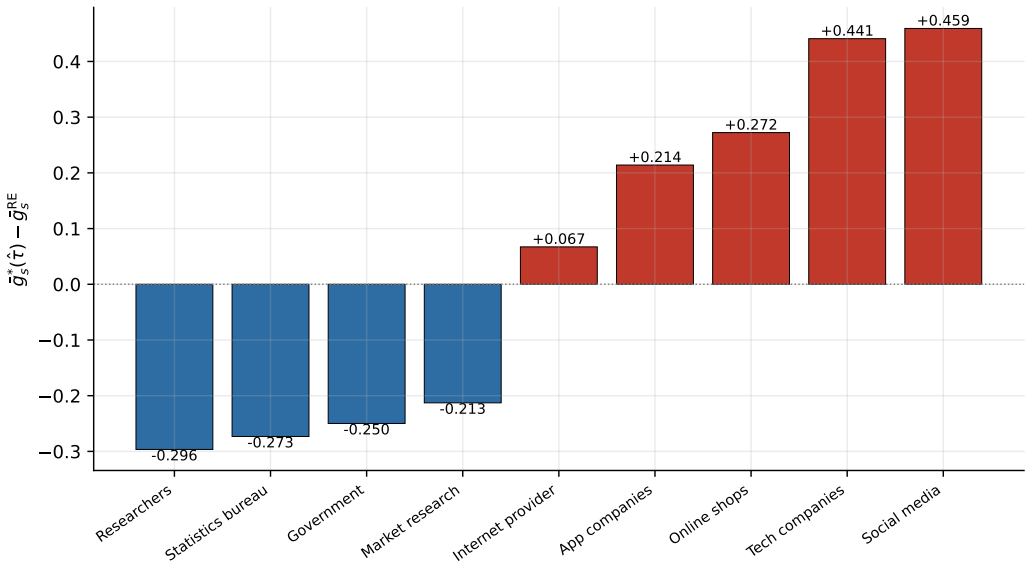


Figure 6. The naive subsidy. Bars report $D_s \equiv \bar{g}_s^*(\hat{\tau}) - \bar{g}_s^{\mathrm{RE}}$, the part of each organization's pool average that exists only because the population is boundedly rational. Red bars are organizations with $\hat{\psi}_s > 0$, blue bars $\hat{\psi}_s < 0$. The gap follows the strategic-use coefficient $\hat{\psi}_s$, not the trust ranking: market research firms, sixth of nine in mean trust, have a negative gap because $\hat{\psi}_s < 0$. Bootstrap confidence intervals are in Table 17.

The distance between the path at $\hat{\tau}$ and the benchmark is $D_s \equiv \bar{g}_s^*(\hat{\tau}) - \bar{g}_s^{\mathrm{RE}}$, the part of the pool average that exists only because part of the population does not account for selection. We call D_s the *naive subsidy*. Figure 6 reports it. The subsidy is largest at the organizations with the strongest adverse selection: +0.44 and +0.46 standard deviations of income at technology companies and social media, with 95 percent confidence intervals that exclude +0.30. It is negative at the researchers (−0.30), the statistics bureau (−0.27) and the government (−0.25): bounded rationality also mutes the advantageous selection that sophistication would produce at these organizations. Table 20 in Appendix F reports the pools by organization.

8.2 Welfare Effect

Following Bernheim and Rangel (2009), we evaluate welfare at the realized payoff η_{is}^* in equation (8) rather than at the payoff $\eta_{is}^{(k)}$ that the consumer perceives at her cognitive level.

Welfare is the expected realized utility of equation (19), measured in log-odds units of the sharing decision. A level-0 consumer who shares without accounting for $\psi_s(g_i - \bar{g}_s^*)$ still bears it, so her welfare at organization s is her sharing probability times the realized value V_{is}^*, plus the expected taste shock $H(q)$. For each respondent we integrate over her posterior distribution of cognitive level and privacy cost given her trust ratings and donation decision.

Table 10 reports the result. Moving from $\hat{\tau} = 0.32$ to $\tau_{\text{cf}} = 3$ lowers mean welfare by 0.087, or 4.3 percent of baseline welfare (95 percent confidence interval 2.7 to 7.1 percent). Only 11 percent of respondents gain. Table 17 reports bootstrap confidence intervals for all counterfactual statistics.

Table 10. Realized welfare under increasing sophistication

Mean welfare, baseline ($\hat{\tau} = 0.32$)	2.040
Mean welfare, counterfactual ($\tau_{\text{cf}} = 3$)	1.953
Δ mean welfare	-0.087
decision channel	+0.092
equilibrium channel	-0.179
Δ welfare, 10th / 90th percentile	-0.207 / +0.014
Share of consumers better off (%)	11.1
Baseline decision loss	0.113
Share of decision loss corrected (%)	81.0

Notes. Welfare is expected realized utility, equation (19), in log-odds units of the sharing decision, summed over organizations and averaged over respondents, with each respondent integrated over her posterior distribution of cognitive level and privacy cost. Decision and equilibrium channels are defined in equation (23). The baseline decision loss is the welfare a respondent loses relative to choosing with correct beliefs about the baseline market.

8.3 Decomposition

Let $U_i(q, V^*)$ denote respondent i's welfare when she shares with probabilities q and receives realized values V^*. We decompose the change in welfare as

$$\Delta U_i = \underbrace{U_i(q^{\text{cf}}, V^{*,0}) - U_i(q^0, V^{*,0})}_{\text{decision channel}} + \underbrace{U_i(q^{\text{cf}}, V^{*,\text{cf}}) - U_i(q^{\text{cf}}, V^{*,0})}_{\text{equilibrium channel}}. \tag{23}$$

The decision channel is the change in welfare from making different decisions, evaluated in the baseline market. It equals the reduction in the respondent's decision loss, the gap between her welfare and the welfare she would obtain by choosing with correct beliefs about that market (Appendix E). The equilibrium channel is the change in welfare from the change in the market itself, evaluated at the new decisions.

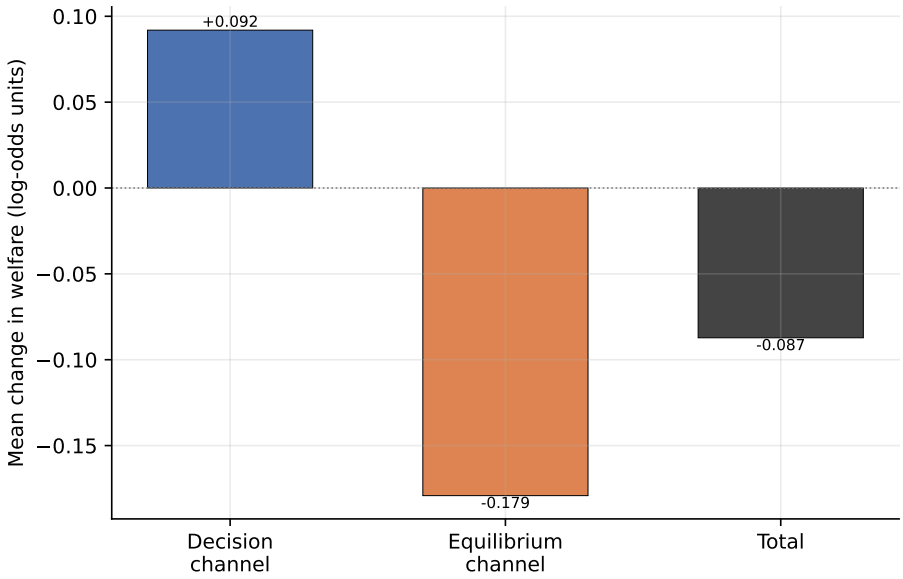


Figure 7. Decomposition of the welfare effect of increasing sophistication.

The decision channel is +0.092 (95 percent confidence interval 0.061 to 0.140): increasing sophistication corrects 81 percent of the baseline decision loss. The equilibrium channel is −0.179 (−0.272 to −0.118), about twice as large, so mean welfare falls (Figure 7).

8.4 The Mechanism: Sorting in the Sharer Pool

Figure 8 plots the realized $\bar{g}_s^*$ at baseline and under increasing sophistication.

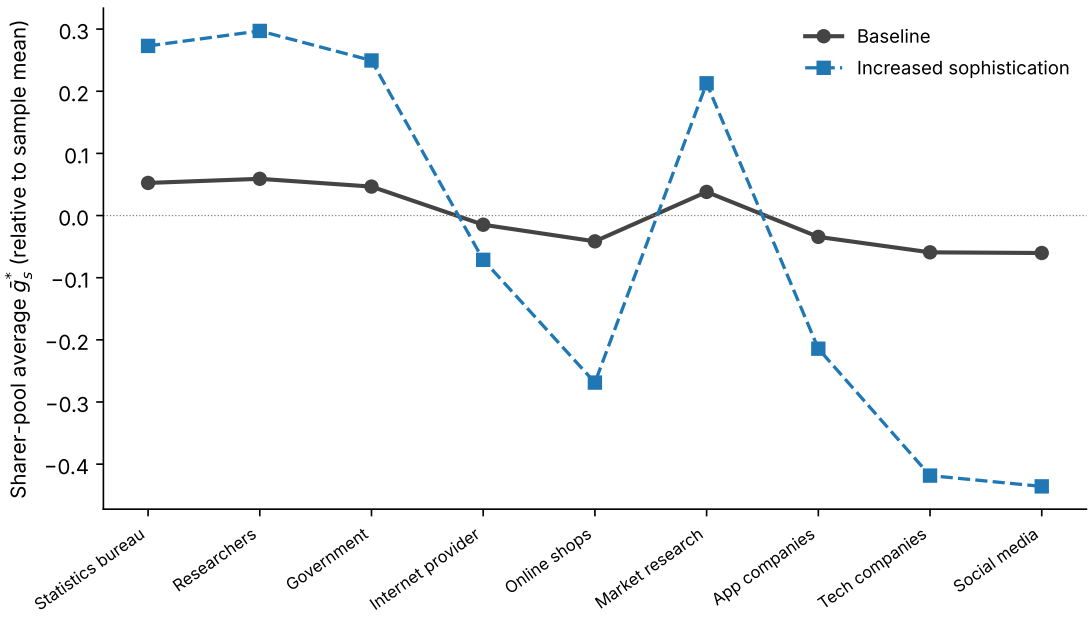


Figure 8. Realized sharer-pool average $\bar{g}_s^*$ at baseline and under increasing sophistication, by organization. High-type consumers leave the organizations with $\hat{\psi}_s > 0$ and join those with $\hat{\psi}_s < 0$.

Sophisticated high-type consumers leave the organizations that take more from higher types and join those that favor them. The pool average falls by 0.36 at technology companies and 0.38 at social media, and rises by 0.24 at the researchers and 0.22 at the statistics bureau. Sharing rates change little; what changes is who shares.

The realized payoff change is

$$\eta_{is}^{*,\mathrm{cf}} - \eta_{is}^{*,0} = \psi_s\big(\bar{g}_s^{*,\mathrm{cf}} - \bar{g}_s^{*,0}\big), \tag{24}$$

in which the consumer's own type cancels. Because pools move in the direction opposite to the sign of ψ_s, the right-hand side is negative at every organization. Sorting is a transfer among members: the consumer who moves to the pool that favors her type gains, which appears in her decision channel, and the remaining members lose, which appears in their equilibrium channel. The equilibrium channel is negative at all nine organizations and largest where $|\psi_s|$ and the change in the pool are largest: the researchers, the statistics bureau,

technology companies, and social media.

8.5 Distributional Effects by Cognitive Level

Table 11 reports the two channels by baseline cognitive level, by transition between levels, and by income decile, and Figure 9 plots them by baseline level.

Table 11. Welfare decomposition by baseline cognitive level, transition, and income decile

Group	Share	Decision	Equilibrium	Total	% better off
All consumers	1.000	+0.092	-0.179	-0.087	11.1
By baseline cognitive level					
Level 0 (naive)	0.728	+0.128	-0.182	-0.054	15.7
Level 1	0.235	-0.005	-0.171	-0.176	0.0
Level 2	0.038	-0.006	-0.171	-0.177	0.0
By transition (baseline → counterfactual level)					
$0 \to 0$	0.118	+0.000	-0.181	-0.181	0.0
$0 \to 1$	0.353	+0.155	-0.190	-0.035	18.3
$0 \to 2$	0.257	+0.150	-0.172	-0.022	20.1
$1 \to 2$	0.235	-0.005	-0.171	-0.176	0.0
$2 \to 2$	0.038	-0.006	-0.171	-0.177	0.0
By income decile					
D1	0.100	+0.583	-0.196	+0.387	88.2
D2	0.100	+0.075	-0.172	-0.096	0.3
D3	0.100	+0.017	-0.166	-0.150	0.0
D4	0.100	-0.000	-0.165	-0.166	0.0
D5	0.100	-0.002	-0.173	-0.175	0.0
D6	0.100	+0.004	-0.176	-0.172	0.0
D7	0.100	+0.013	-0.178	-0.165	0.0
D8	0.100	+0.030	-0.183	-0.153	0.0
D9	0.100	+0.052	-0.184	-0.132	0.0
D10	0.100	+0.145	-0.198	-0.053	22.3

Notes. Respondents are weighted by their posterior probability of each baseline level. Under the counterfactual, a consumer at baseline level k moves to level $k' \geq k$ with the probabilities of the monotone coupling. Channels are defined in equation (23). Income deciles are based on log net monthly income.

Decision gains accrue to consumers who were naive. The decision channel is +0.128 for consumers who were naive at baseline. Their decisions ignored the pool term, so they have the most to correct. For consumers already at level 1 or 2 it is slightly negative (-0.005

and −0.006): their forecasts adjust to the more sophisticated population, which moves their choices away from the best response to the baseline market.

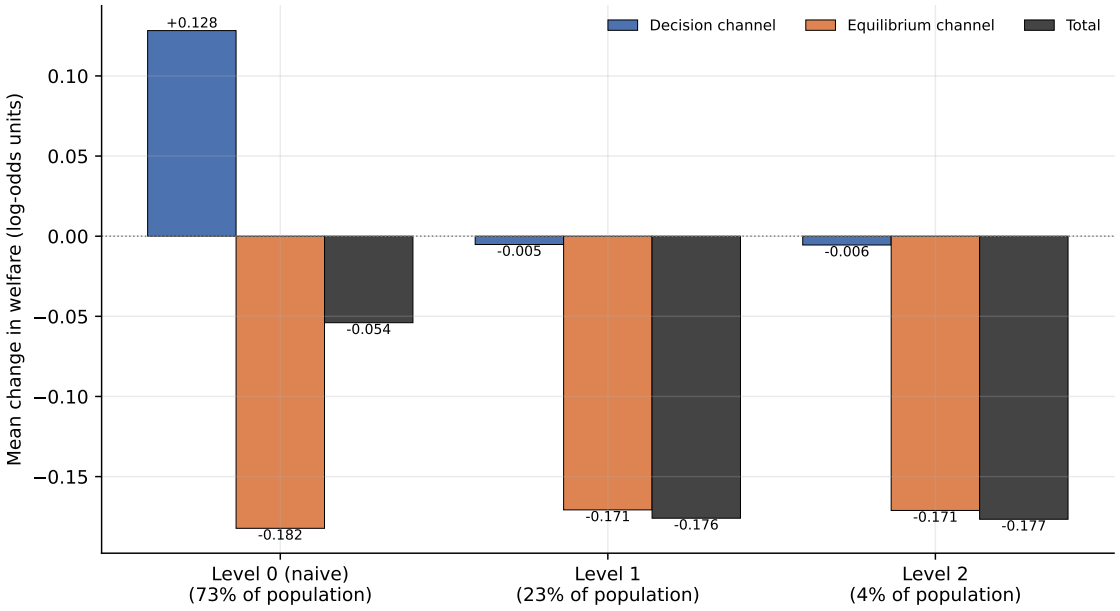


Figure 9. Welfare change under increasing sophistication, by baseline cognitive level.

The externality falls on everyone. The equilibrium channel is between −0.17 and −0.18 for every baseline level. By equation (24), the change in payoffs at an organization depends only on the change in its pool and on ψ_s, not on the consumer's own level; the channel differs across levels only through how much each consumer shares.

Net effects. Consumers who were naive lose 0.054 on average, and 16 percent of them gain. Consumers who were already sophisticated lose 0.176 and none gain. The largest losses fall on consumers who remain naive under the counterfactual (12 percent of the population): they bear the externality (−0.181) without any decision gain.

8.6 Distributional Effects by Income

Figure 10 plots the decomposition by income decile.

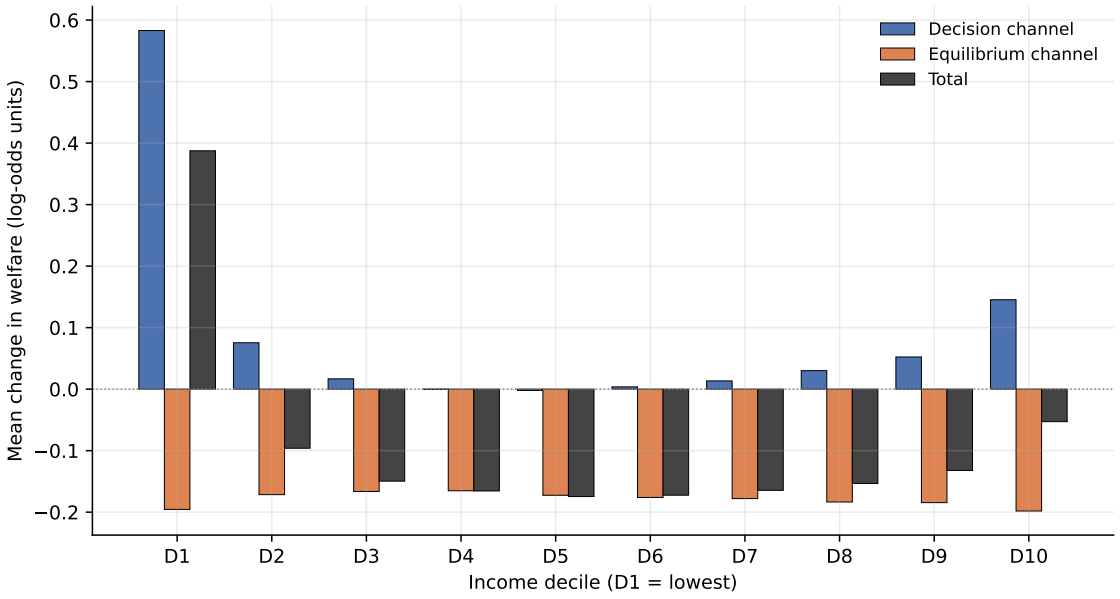


Figure 10. Welfare change under increasing sophistication, by income decile. Decision gains are concentrated in the tails of the income distribution; the equilibrium loss is similar across deciles.

Decision gains are concentrated in the tails of the income distribution. The term a naive consumer ignores, $\psi_s(g_i - \bar{g}_s^*)$, is largest for consumers whose income is far from the pool average, so they have the most to correct. The bottom decile gains on net (+0.39; 88 percent of its members gain), because the distribution of log income has a long left tail and these consumers are furthest from the pool average. The top decile loses little (−0.05). Deciles 3 to 9, which have little to correct and bear the full externality, lose between 0.13 and 0.18. The equilibrium loss is similar across deciles, between 0.16 and 0.20.

8.7 Sophistication That Varies with Characteristics

The main counterfactual uses a common τ, so differences in sophistication across income groups enter only through each respondent's posterior. In the model of Section 7.1, sophistication rises with income, so naive consumers are concentrated at the bottom of the income distribution. We repeat the counterfactual under that model. The policy raises every respondent's cognitive intensity by the same factor, increasing δ_0 so that the share of naive

consumers falls to 12 percent, as in the main counterfactual.

Table 18 compares the results. Mean welfare falls by 3.4 percent, within the confidence interval of the main estimate. The decision and equilibrium channels are about one-fifth smaller, 14 percent of respondents gain, and the threshold λ^* is unchanged at 0.49. The bottom income decile gains in both models; the share of gainers in the top decile rises from 22 to 43 percent (Figure 15 in Appendix F).

8.8 Subjective versus Realized Welfare

An alternative is the subjective, option-value measure $W_i^{\text{subj}} = \sum_s \sum_k \hat{p}_{i,k} \log\left(1 + e^{\gamma_0 + \gamma_1 \eta_{is}^{(k)}}\right)$, the inclusive value of Williams (1977) and Small and Rosen (1981) evaluated at perceived payoffs, where $\hat{p}_{i,k}$ is respondent i's posterior probability of level k. It credits each consumer with the value of her decision as she perceived it, including costs she did not anticipate.

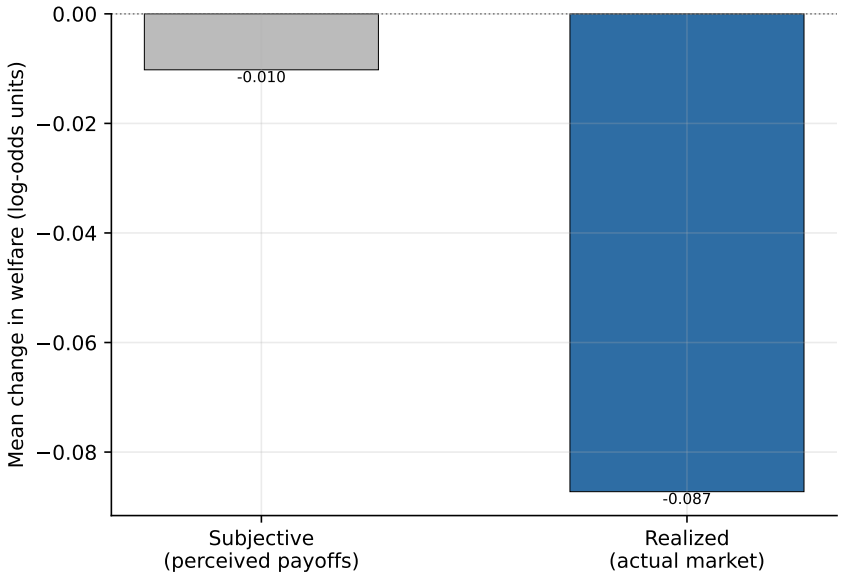


Figure 11. Subjective (option-value) versus realized welfare effect of increasing sophistication.

Measured at perceived payoffs, increasing sophistication appears nearly neutral (-0.010). The realized loss is more than eight times larger (Figure 11). The subjective measure scores choices at consumers' beliefs and so misses most of the change in realized payoffs that sorting

produces. Consistent with Bernheim and Rangel (2009), we use the realized measure for our main results.

8.9 Welfare Relevance of Sorting

Trust ratings and the data donation decision identify ψ_s, the slope of the payoff in a consumer's own type relative to the pool. They do not identify how much of this slope is redistribution among members through the pooled benefit and how much reflects real differences in the surplus that a member's data create, for example through better matching. Our main specification treats it as redistribution. Under break-even, the transfer terms sum to zero, so the efficient allocation is the one in which consumers decide on $v_s - \theta_i$, which is how naive consumers decide. Sophistication then adds only a private incentive to sort, and information that lets consumers sort has private but no social value (Hirshleifer, 1971).

Following Handel and Kolstad (2015), who report welfare with and without the welfare relevance of a component of utility that choice data cannot classify, we keep behavior at the estimated model and vary only the welfare-relevant payoff. Let $\lambda \in [0, 1]$ be the share of the personalization ψ_s that reflects real surplus. The welfare-relevant payoff in a counterfactual with realized pool $\bar{g}_s^*$ is then

$$\eta_{is}^*(\lambda) = v_s - \theta_i - \psi_s g_i + \psi_s \left[(1 - \lambda)\, \bar{g}_s^* + \lambda\, \bar{g}_s^{*,0} \right], \tag{25}$$

where $\bar{g}_s^{*,0}$ is the baseline pool, at which v_s is estimated. At the baseline, $\eta_{is}^*(\lambda) = \eta_{is}^*$ for every λ, so the estimates and baseline welfare are unchanged. When $\lambda = 1$, a member's payoff depends only on her own type and there is no pooling externality. For each respondent, $\Delta U_i(\lambda)$ equals the decision channel plus $(1 - \lambda)$ times the equilibrium channel.

Table 12. Welfare effect of increasing sophistication by welfare relevance of sorting

	Increase in sophistication ($\hat{\tau} \to \tau_{\text{cf}}$)					Relative to baseline	
λ	ΔW	%	Decision	Equilibrium	% gain	All naive	RE
0.00	-0.087	-4.3	+0.092	-0.179	11.1	+0.020	-0.102
0.25	-0.042	-2.1	+0.092	-0.134	15.3	+0.010	-0.049
0.50	+0.002	+0.1	+0.092	-0.090	21.5	+0.000	+0.005
0.75	+0.047	+2.3	+0.092	-0.045	33.2	-0.009	+0.059
1.00	+0.092	+4.5	+0.092	-0.000	73.1	-0.019	+0.113
Sign change of ΔW at $\lambda^* = 0.49$							

Notes. λ is the share of the type slope ψ_s that reflects real surplus; $1 - \lambda$ is redistribution among members. Behavior is the estimated model in every row. "All naive" and "RE" report welfare when all consumers are level 0 and when consumers best-respond to the welfare-relevant payoff, relative to baseline welfare.

Table 12 and Figure 12 report the results. Mean welfare falls with sophistication when $\lambda = 0$ and rises when $\lambda = 1$ (+0.092, or 4.5 percent). The sign changes at $\lambda^* = 0.49$ (95 percent confidence interval 0.466 to 0.495), and the organization-specific thresholds lie between 0.44 and 0.54 (Table 19). A threshold near one half is to be expected. For a small change, a consumer who sorts gains roughly half of the transfer she shifts, the usual triangle, while the other members lose all of it. Sophistication therefore lowers welfare whenever more than about half of the slope ψ_s is redistribution. Our estimates determine the size and distribution of the effects and place the threshold at 0.49.

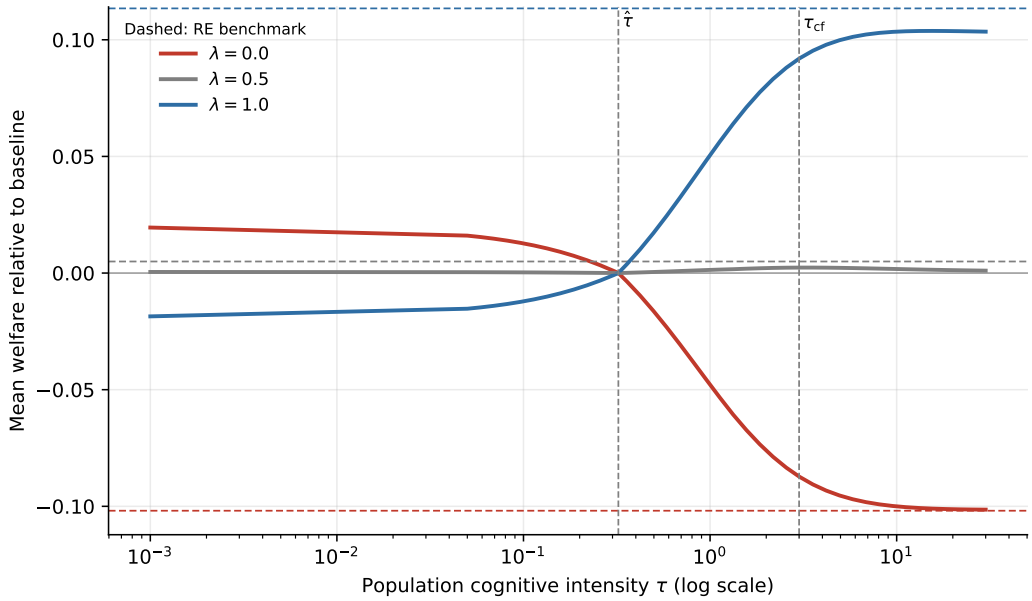


Figure 12. Mean welfare along the sophistication path, relative to baseline, for $\lambda = 0$, 0.5 and 1. Dashed lines are the corresponding rational-expectations benchmarks.

Discussion

Increasing sophistication corrects most of consumers' decision errors but lowers average welfare by 4.3 percent, because the sorting it induces is a transfer among members. Consumers who were naive gain from better decisions, while every consumer bears the deterioration of the pools in which she remains. The net effect is negative for most consumers, and largest for those who were already sophisticated and for those who stay naive.

The result has a direct interpretation in Akerlof (1970)'s framework. Sophistication makes consumers act on the selection in the data market, and high types withdraw from the organizations that take more from them and move to those that favor them. The pools move in the direction the lemons literature predicts, but not all the way to the rational-expectations benchmark: at $\tau_{\text{cf}} = 3$ they close about four-fifths of the naive subsidy of Figure 6. How far they move is governed by the estimated distribution of cognitive levels and by the dispersion of θ_i, which sustains some sharing even where $|\psi_s|$ is large.

These results bear on policies that require organizations to disclose how they use consumer data. The usual justification is consumer protection: consumers misperceive how their data will be used, and disclosure closes the gap. We find that closing the gap improves individual decisions but, in a pooled market, mainly redistributes surplus among consumers. Whether such a policy raises welfare depends on whether the sorting it induces creates real surplus, which requires a share of at least about one half in our estimates, or merely moves transfers between consumers.

9 Conclusion

We study consumer data sharing as a selection market in which consumers differ in how far they reason about who else shares. Using trust ratings for nine types of organizations and a real data-donation decision from 3,132 respondents in the LISS panel, we estimate a cognitive-hierarchy model in which the terms a sharer receives depend on the average type of the pool. We find that 73 percent of consumers decide without reasoning about the pool, and that consumers expect public and research organizations to favor higher-income sharers and most commercial organizations to take more from them.

Bounded rationality shapes how much selection occurs. At the estimated sophistication, pools at most commercial organizations are much less adversely selected than they would be under rational expectations, and pools at public and research organizations are much less advantageously selected. A policy that raises sophistication moves pools most of the way toward the rational-expectations benchmark. It corrects 81 percent of consumers' decision errors but lowers average welfare by 4.3 percent, and only 11 percent of consumers gain, all of them initially naive. The loss arises because, in a pooled market, the sorting that sophistication induces mainly moves transfers between consumers.

These conclusions depend on what the personalization we estimate represents. If more than about half of the type gradient reflects real differences in the surplus that consumers'

data create rather than redistribution, raising sophistication increases welfare. Our data identify how consumers perceive the consequences of sharing, not organizations' costs or the source of these differences. Combining consumer-side estimates with data on organizations' pricing and service decisions, and letting organizations adjust their use of data to the pool they attract, are natural next steps.

References

Alessandro Acquisti and Hal R. Varian. Conditioning prices on purchase history. *Marketing Science*, 24(3):367–381, 2005.

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–492, 2016.

Sumit Agarwal, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel. Regulating consumer financial products: Evidence from credit cards. *The Quarterly Journal of Economics*, 130(1):111–164, 2015.

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

S Nageeb Ali, Greg Lewis, and Shoshana Vasserman. Voluntary disclosure and personalized pricing. *The Review of Economic Studies*, 90(2):538–571, 2023.

Hunt Allcott. The welfare effects of misperceived product costs: Data and calibrations from the automobile market. *American Economic Journal: Economic Policy*, 5(3):30–66, 2013.

Hunt Allcott, Benjamin B. Lockwood, and Dmitry Taubinsky. Regressive sin taxes, with an application to the optimal soda tax. *Quarterly Journal of Economics*, 134(3):1557–1626, 2019.

Mark Armstrong. Competition in two-sided markets. *RAND Journal of Economics*, 37(3): 668–691, 2006.

Susan Athey, Christian Catalini, and Catherine E. Tucker. The digital privacy paradox: Small money, small costs, small talk. NBER Working Paper 23488, National Bureau of Economic Research, 2017.

Dimitris Batzilis, Sonia Jaffe, Steven Levitt, John A. List, and Jeffrey Picel. Behavior in strategic settings: Evidence from a million rock-paper-scissors games. *Games*, 10(2):18, 2019. doi: 10.3390/g10020018.

Volker Benndorf, Dorothea Kübler, and Hans-Theo Normann. Privacy concerns, voluntary disclosure of information, and unraveling: An experiment. *European Economic Review*, 75:43–59, 2015.

Dirk Bergemann, Alessandro Bonatti, and Tan Gan. The economics of social data. *RAND Journal of Economics*, 53(2):263–296, 2022.

B. Douglas Bernheim and Antonio Rangel. Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124(1): 51–104, 2009.

Alessandro Bonatti and Gonzalo Cisternas. Consumer scores and price discrimination. *The Review of Economic Studies*, 87(2):750–791, 2020.

Alexander L. Brown, Colin F. Camerer, and Dan Lovallo. To review or not to review? Limited strategic thinking at the movie box office. *American Economic Journal: Microeconomics*, 4(2):1–26, 2012.

Alexander L. Brown, Colin F. Camerer, and Dan Lovallo. Estimating structural models of equilibrium and cognitive hierarchy thinking in the field: The case of withheld movie critic reviews. *Management Science*, 59(3):733–747, 2013. doi: 10.1287/mnsc.1120.1563.

Jennifer Brown, Tanjim Hossain, and John Morgan. Shrouded attributes and information suppression: Evidence from the field. *The Quarterly Journal of Economics*, 125(2):859–876, 2010.

Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898, 2004.

CentERdata. LISS panel: Longitudinal internet studies for the social sciences, 2024. Tilburg University, https://www.dataarchive.lissdata.nl.

Jay Pil Choi, Doh-Shin Jeon, and Byung-Cheol Kim. Privacy and personal data collection with information externalities. *Journal of Public Economics*, 173:113–124, 2019.

Vincent Conitzer, Curtis R Taylor, and Liad Wagman. Hide and seek: Costly consumer privacy in a market with repeat purchases. *Marketing Science*, 31(2):277–292, 2012.

Rense Corten, Laura Boeschoten, Thijs Carrière, Stein Jongerius, Bella Struminskaya, Joris Mulder, Parisa Zahedi, Shiva Nadi Najafabadi, and Adriënne Mendrik. Assessing mobile instant messenger networks with donated data. *Social Network Analysis and Mining*, 2025. doi: 10.1007/s13278-025-01550-8.

Miguel A. Costa-Gomes and Vincent P. Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768, 2006.

Vincent P. Crawford and Nagore Iriberri. Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770, 2007.

Vincent P. Crawford, Miguel A. Costa-Gomes, and Nagore Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62, 2013.

C. Mitchell Dayton and George B. Macready. Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178, 1988.

Jean-Pierre Dubé and Sanjog Misra. Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1):131–189, 2023.

Liran Einav and Amy Finkelstein. Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic perspectives*, 25(1):115–138, 2011.

Liran Einav, Amy Finkelstein, and Mark R Cullen. Estimating welfare in insurance markets using variation in prices. *The Quarterly Journal of Economics*, 125(3):877–921, 2010.

Xavier Gabaix. Behavioral inattention. *Handbook of Behavioral Economics*, 2:261–343, 2019.

Xavier Gabaix and David Laibson. Shrouded attributes, consumer myopia, and information suppression in competitive markets. *The Quarterly Journal of Economics*, 121(2):505–540, 2006.

David Gill and Victoria Prowse. Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, 124(6):1619–1676, 2016.

Avi Goldfarb and Verina F. Que. The economics of digital privacy. *Annual Review of Economics*, 15:267–286, 2023.

Avi Goldfarb and Catherine E. Tucker. Privacy regulation and online advertising. *Management Science*, 57(1):57–71, 2011.

Avi Goldfarb and Mo Xiao. Who thinks about the competition? Managerial ability and strategic entry in US local telephone markets. *American Economic Review*, 101(7):3130–3161, 2011. doi: 10.1257/aer.101.7.3130.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461–483, 1981.

Michael D Grubb and Matthew Osborne. Cellular service demand: Biased beliefs, learning, and bill shock. *American Economic Review*, 105(1):234–271, 2015.

Benjamin R Handel. Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, 103(7):2643–2682, 2013.

Benjamin R. Handel and Jonathan T. Kolstad. Health insurance for "humans": Information frictions, plan choice, and consumer welfare. *American Economic Review*, 105(8):2449–2500, 2015.

Benjamin R Handel, Jonathan T Kolstad, and Johannes Spinnewijn. Information frictions and adverse selection: Policy interventions in health insurance markets. *Review of Economics and Statistics*, 101(2):326–340, 2019.

Paul Heidhues and Botond Kőszegi. Behavioral industrial organization. *Handbook of Behavioral Economics*, 1:517–612, 2018.

Benjamin E. Hermalin and Michael L. Katz. Privacy, property rights and efficiency: The economics of privacy as secrecy. *Quantitative Marketing and Economics*, 4(3):209–239, 2006.

Jack Hirshleifer. The private and social value of information and the reward to inventive activity. *American Economic Review*, 61(4):561–574, 1971.

Ali Hortaçsu, Fernando Luco, Steven L. Puller, and Dongni Zhu. Does strategic ability affect efficiency? Evidence from electricity markets. *American Economic Review*, 109(12): 4302–4342, 2019. doi: 10.1257/aer.20172015.

Shota Ichihashi. Online privacy and information disclosure by consumers. *American Economic Review*, 110(2):569–595, 2020.

Annette Jäckle, Jonathan Burton, and Mick P. Couper. Understanding Society: Minimising selection biases in data collection using mobile apps. *Fiscal Studies*, 44(4):361–376, 2023.

Daniel Kahneman, Peter P Wakker, and Rakesh Sarin. Back to bentham? explorations of experienced utility. *The quarterly journal of economics*, 112(2):375–406, 1997.

Michael Kummer and Patrick Schulte. When private information settles the bill: Money and privacy in Google's market for smartphone applications. *Management Science*, 65(8): 3470–3494, 2019.

Christopher G. Leggett. Environmental valuation with imperfect information: The case of the random utility model. *Environmental and Resource Economics*, 23(3):343–355, 2002.

Tesary Lin. Valuing intrinsic and instrumental preferences for privacy. *Marketing Science*, 41(4):663–681, 2022.

Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391, 1981.

Rodrigo Montes, Wilfried Sand-Zantman, and Tommaso Valletti. The value of personal information in online markets with endogenous privacy. *Management Science*, 65(3):1342–1362, 2019.

Sendhil Mullainathan, Joshua Schwartzstein, and William J. Congdon. A reduced-form approach to behavioral public finance. *Annual Review of Economics*, 4:511–540, 2012.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326, 1995.

Robert Östling, Joseph Tao-yi Wang, Eileen Y. Chou, and Colin F. Camerer. Testing game theory in the field: Swedish LUPI lottery games. *American Economic Journal: Microeconomics*, 3(3):1–33, 2011.

Leslie E. Papke and Jeffrey M. Wooldridge. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6):619–632, 1996.

Richard A. Posner. The economics of privacy. *American Economic Review*, 71(2):405–409, 1981.

Verina F. Que. How do past privacy choices shape the future? Working paper, available at SSRN: https://ssrn.com/abstract=6432042., 2026.

Kenneth A. Small and Harvey S. Rosen. Applied welfare economics with discrete choice models. *Econometrica*, 49(1):105–130, 1981.

Michael Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973.

Johannes Spinnewijn. Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association*, 13(1):130–167, 2015.

Dale O. Stahl and Paul W. Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

George J. Stigler. The economics of information. *Journal of Political Economy*, 69(3): 213–225, 1961.

Bella Struminskaya, Peter Lugtig, Vera Toepoel, Barry Schouten, Deirdre Giesen, and Ralph Dolmans. Sharing data collected with smartphone sensors: Willingness, participation, and nonparticipation bias. *Public Opinion Quarterly*, 85(S1):423–462, 2021.

Curtis R Taylor. Consumer privacy and the market for customer information. *RAND Journal of Economics*, 35(4):631–650, 2004.

Kenneth Train. Welfare calculations in discrete choice models when anticipated and experienced attributes differ: A guide with examples. *Journal of Choice Modelling*, 16:15–22, 2015.

Catherine E. Tucker. Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, 51(5):546–562, 2014.

André Veiga and E Glen Weyl. Product design in selection markets. *The Quarterly Journal of Economics*, 131(2):1007–1056, 2016.

H. C. W. L. Williams. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A*, 9(3):285–344, 1977.

A Additional Tables and Figures

Table 13. Mean trust ratings by organization, full sample.

Organization	Mean	SD
Statistics bureau	3.03	0.77
Researchers	3.02	0.76
Government	2.92	0.78
Internet provider	2.71	0.78
Online shops	2.46	0.81
Market research	2.37	0.84
App companies	2.27	0.75
Tech companies	2.21	0.83
Social media	1.98	0.81

Notes. Mean trust on the four-point scale, all 3,598 respondents who reached the trust module, of whom 3,588 rated at least one organization. The ordering is the same as in the analysis sample (Table 1), on which the harm rank is defined.

B Functional Form of the Trust–Harm Gradient

Equations (1) and (2) in Section 3 treat institutional harm as linear in the 1–9 rank. Table 14 re-estimates the joint specification under alternative functional forms for harm: a binary high/low-harm indicator, splitting either between the public/research cluster and all commercial settings (ranks 4–9) or at the median rank (ranks 6–9), and a three-cluster specification: public and research organizations (ranks 1–3), intermediate commercial organizations (ranks 4–6), and technology platforms (ranks 7–9).

The heterogeneity patterns are not an artifact of the linear-in-rank choice. The comprehension interaction is negative and statistically significant under every discretization, with magnitudes consistent with the linear model: the linear estimates imply a high-vs.-low comprehension-interaction contrast of roughly -0.032 for the ranks-4–9 split, close to the estimated -0.039, and the three-cluster interactions are monotone in harm (-0.029 for the intermediate cluster and -0.049 for the technology cluster, relative to the public cluster). The high-education interaction behaves the same way, while the digital-literacy interaction is small and never significant at the five-percent level under any functional form, mirroring its behavior in the joint linear specification. A fully non-parametric specification, with eight rank dummies each interacted with comprehension, rejects joint equality of the interaction terms ($p < 0.001$) and reproduces the monotone step-down (see the table notes).

C Trust Pass-Through and the Bias Decomposition

Table 15 reports the full estimates of the linear-probability model in equation (4). Column (1) reports the unconditional trust pass-through ($\hat{\beta} = 0.286$); columns (2)–(4) successively add demographic characteristics, survey measures of comprehension and skill, and WhatsApp usage. The pass-through declines from 0.286 in column (1) to 0.228 in column (2) to approximately 0.13 in columns (3) and (4), confirming that a substantial portion of the raw trust–sharing correlation reflects co-variation between trust and the proxy characteristics rather than a direct trust channel. The residual variation in actual sharing not captured by stated trust, the person-specific bias, has a standard deviation of approximately 0.11 across respondents and a 5–95 percentile range of roughly -0.02 to $+0.32$ in column (2). Age^2, the survey-wave indicator, and the image-treatment indicator are excluded from the specification; education enters as a single continuous variable in line with Table 3.

Table 14. Robustness of the Within-Person Trust–Harm Gradient to Alternative Functional Forms for Harm

	(1) Linear rank (1–9)	(2) High/low (ranks 4–9)	(3) High/low (ranks 6–9)	(4) Three clusters
Dependent variable: Trust in organization (y_{is}, 1–4)				
Harm rank (centered)	-0.143***	—	—	—
	(0.002)			
High harm (indicator)	—	-0.682***	-0.640***	—
		(0.011)	(0.009)	
Mid-harm cluster	—	—	—	-0.495***
				(0.010)
High-harm cluster	—	—	—	-0.869***
				(0.013)
Harm × Comprehension				
Harm measure × Compr.	-0.007***	-0.039***	-0.024***	—
	(0.001)	(0.005)	(0.004)	
Mid cluster × Compr.	—	—	—	-0.029***
				(0.005)
High cluster × Compr.	—	—	—	-0.049***
				(0.006)
Harm × Digital literacy				
Harm measure × Dig. lit.	-0.002	-0.000	0.002	—
	(0.002)	(0.012)	(0.010)	
Mid cluster × Dig. lit.	—	—	—	0.020*
				(0.011)
High cluster × Dig. lit.	—	—	—	-0.021
				(0.014)
Harm × High education				
Harm measure × High educ.	-0.062***	-0.347***	-0.185***	—
	(0.004)	(0.023)	(0.018)	
Mid cluster × High educ.	—	—	—	-0.259***
				(0.021)
High cluster × High educ.	—	—	—	-0.434***
				(0.028)
Person fixed effects	Yes	Yes	Yes	Yes
Observations	28,044	28,044	28,044	28,044
R^2 (within)	0.315	0.243	0.225	0.299

Notes: Each column re-estimates the joint specification of equation (2) (all three moderators) under an alternative functional form for institutional harm. Column (1) uses the linear 1–9 harm rank. Column (2) discretizes harm into a binary indicator separating the public-sector/research cluster (ranks 1–3) from all commercial settings (ranks 4–9); column (3) uses an above-median split (ranks 6–9 vs. 1–5). Column (4) uses three institutional clusters: public/research (ranks 1–3, omitted), intermediate commercial (ranks 4–6), and technology platforms (ranks 7–9). All harm measures and moderators are centered at their sample means; standard errors are clustered at the person level. The magnitudes are mutually consistent: the linear model implies a high-vs.-low contrast of $-0.143 \times 4.5 \approx -0.64$ for the column-(2) split, close to the estimated -0.68, and the cluster interactions with comprehension (-0.029 mid, -0.049 high) bracket the linear implication at the cluster distances. A fully non-parametric specification with eight rank dummies and rank-dummy × comprehension interactions yields an almost perfectly monotonic step-down in the dummies and rejects joint equality of the interactions (χ^2 Wald $p < 0.001$). Digital literacy is on the mean-of-items (1–5) scale. The estimation sample drops the 16 respondents with missing digital literacy (3,116 respondents); the comprehension moderator is mean-imputed for the 378 respondents with a missing quiz score. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 15. From Self-Reported Trust to Actual Sharing

	(1) Trust only	(2) Trust + demos	(3) Trust + demos + skills	(4) Full controls
Dependent variable: Actual sharing $Y_{share} \in \{0,1\}$ (donation to researchers)				
Trust (researchers, normalized to [0,1])	0.286***	0.228***	0.129***	0.130***
	(0.036)	(0.035)	(0.034)	(0.034)
Female		-0.035*	-0.034**	-0.029
		(0.018)	(0.017)	(0.018)
Age		-0.006***	-0.003***	-0.003***
		(0.001)	(0.001)	(0.001)
Education		0.011*	-0.000	-0.000
		(0.007)	(0.006)	(0.006)
Has partner		-0.015	-0.022	-0.021
		(0.019)	(0.018)	(0.018)
Number of children		0.004	0.003	0.004
		(0.010)	(0.010)	(0.010)
Urbanization		0.001	-0.000	-0.000
		(0.006)	(0.006)	(0.006)
Log net income		0.012	-0.010	-0.009
		(0.016)	(0.016)	(0.016)
Comprehension score			0.061***	0.061***
			(0.003)	(0.003)
Digital literacy			-0.008	-0.008
			(0.010)	(0.010)
Smartphone proficiency			0.048***	0.049***
			(0.011)	(0.011)
WhatsApp frequency				0.025**
				(0.010)
Log (contacts +1)				0.010
				(0.008)
iPhone				0.008
				(0.016)
Constant	0.099***	0.372***	0.101	0.017
	(0.025)	(0.125)	(0.128)	(0.132)
Observations	2,732	2,732	2,732	2,732
R^2	0.023	0.079	0.182	0.184

Notes: Linear probability model. Dependent variable is the binary actual-sharing decision $Y_{\text{share}} \in \{0,1\}$ for the donation request to scientific researchers, the only context with observed sharing behavior. Trust is normalized trust in researchers, mapped from the 1–4 Likert scale to $[0,1]$ via $(T-1)/3$. The coefficient on Trust is the trust pass-through: the increase in the probability of actually sharing when normalized trust moves from 0 to 1. Education enters as a single continuous variable (CBS categories) in line with Table 3; Age^2, the survey-wave indicator, and the image-treatment indicator are excluded. Digital literacy is the mean of five items (1–5), mean-imputed for three respondents who answered none; WhatsApp frequency is coded 1 = daily to 7 = never. The sample consists of the respondents with non-missing values of all controls. Privacy concern is omitted from all specifications; privacy preferences are recovered structurally in Section 4. HC1 robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

D Additional Post-Estimation Results

This appendix collects supporting material for Section 7: model-free gradients of the two posterior objects, and a decomposition of the cognitive-level posterior into its two cumulative margins.

D.1 Model-Free Gradients of the Two Posterior Objects

Figure 13 plots mean $\bar{K}_i$ by income decile and by education category, with no controls. Income matters mostly at the bottom: mean $\bar{K}_i$ is 0.07 in the lowest decile, reaches 0.38 by the third decile, and is roughly flat after that. Education matters only at the tertiary margin: the first four categories are flat, and categories 5 and 6 rise to about 0.34 and 0.40 from a base of about 0.27. Sophistication varies where resources are scarce, not smoothly across the whole distribution.

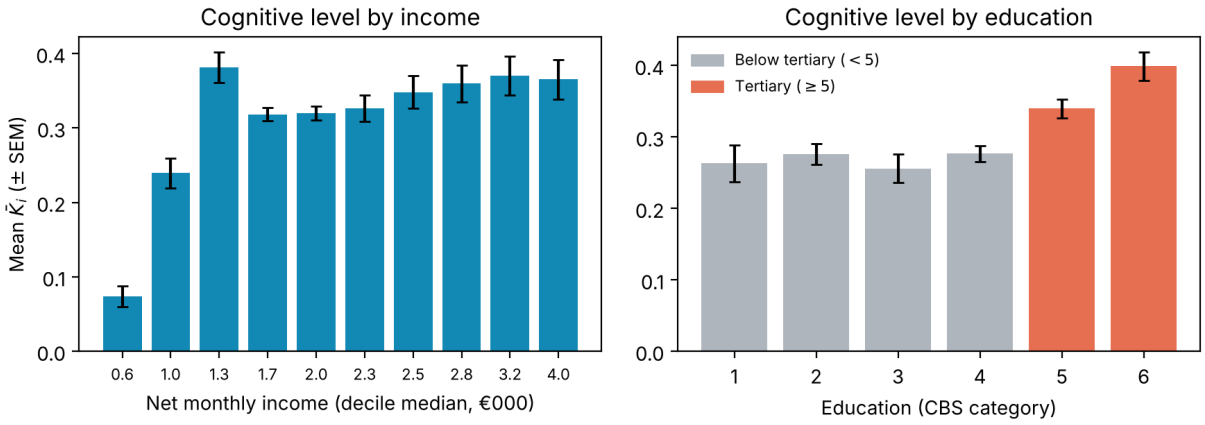


Figure 13. Mean posterior cognitive level $\bar{K}_i$ (± SEM), no controls. *Left:* by net-monthly-income decile (horizontal axis: decile-median income, in thousands of euros). *Right:* by CBS education category, with the tertiary categories (≥5) highlighted.

Figure 14 repeats the same cuts for $\hat{\theta}_i$. Mean $\hat{\theta}_i$ is flat across income deciles and flat across education categories—the two variables that predict $\bar{K}_i$. It declines along the digital-literacy and smartphone-skill ladders: digitally skilled respondents are more willing to share.

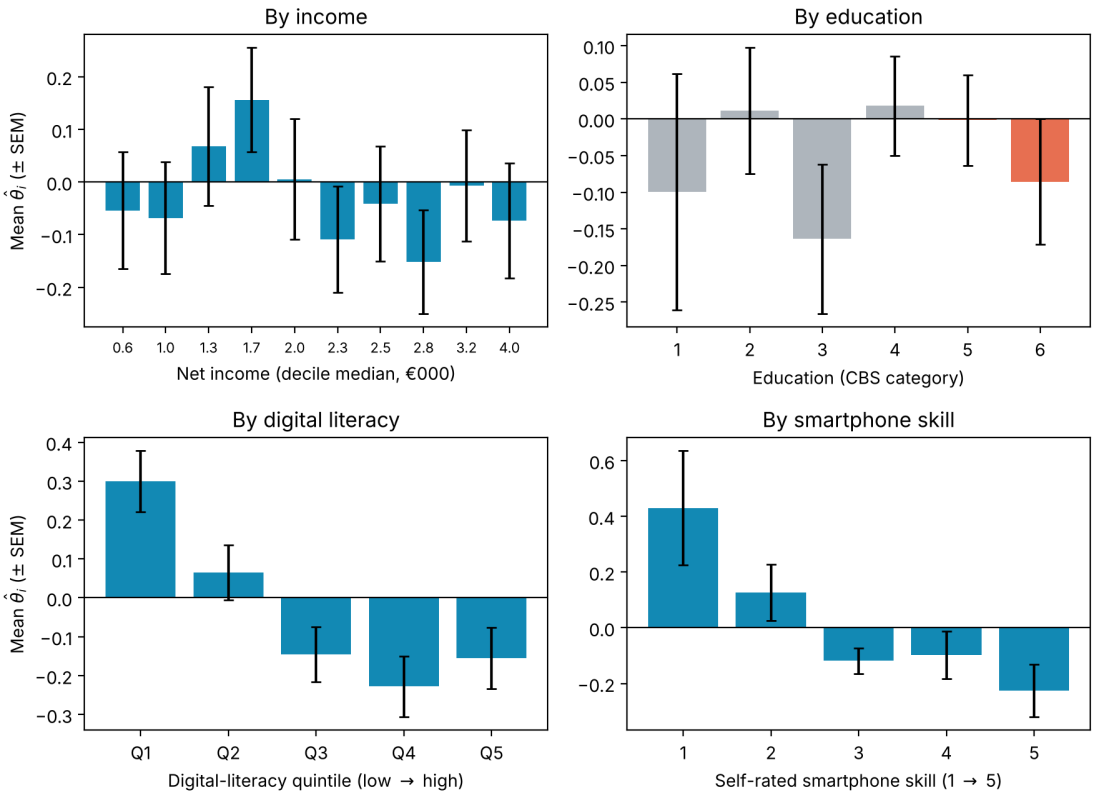


Figure 14. Mean $\hat{\theta}_i$ (± SEM), no controls, in the same layout as Figure 13. *Top:* by net-income decile (in thousands of euros) and by CBS education category (tertiary ≥5 highlighted). *Bottom:* by digital-literacy quintile and by self-rated smartphone skill. Horizontal line: $\hat{\theta}_i = 0$.

D.2 Level Posterior

The mean $\bar{K}_i$ compresses three probabilities into one number, so it cannot say *where* in the level distribution a covariate acts. A covariate could move people out of level 0 without moving anyone into level 2, or the reverse. The distinction matters for the policy experiment in Section 8, which shifts mass out of level 0. We therefore use two cumulative probabilities as outcomes:

$$\Pr(K_i \geq 1) = \hat{p}_{i,1} + \hat{p}_{i,2} \qquad \text{and} \qquad \Pr(K_i \geq 2) = \hat{p}_{i,2}.$$

The first is the probability of doing any strategic reasoning; the second is the probability of reaching the top level. Their sample means, 0.273 and 0.038, are close to the estimated level distribution: 72.8% at level 0, 23.5% at level 1, and 3.8% at level 2. Because both

outcomes are probabilities, Table 16 fits each one twice: by linear-probability OLS and by a fractional logit (Papke and Wooldridge, 1996), reported as average marginal effects. The two estimators agree, so the linear coefficients can be read directly. Continuous regressors are standardized in this table.

Three results. First, income works almost entirely at the first threshold. One standard deviation of log income raises the probability of any strategic reasoning by 6.1 percentage points on a base of 27.3, but the probability of reaching level 2 by only 0.8 points on a base of 3.8. Tertiary education looks the same (+5.4 points at the first threshold, +0.7 at the second). Richer and more educated respondents stop being naive; they do not become highly sophisticated. Second, stated privacy concern predicts neither threshold, consistent with its insignificance in Table 8. Third, age follows an inverted U that peaks at about 48, which is why the specification includes the quadratic.

The first result locates the naive mass in the population: level-0 respondents sit disproportionately at the bottom of the income distribution. Section 8 returns to this fact when it evaluates the counterfactual.

Table 16. Cumulative thresholds of the cognitive-level posterior: linear probability and fractional logit.

	Linear probability (OLS)		Fractional logit (AME)	
	(1) $\Pr(K_i \geq 1)$	(2) $\Pr(K_i \geq 2)$	(3) $\Pr(K_i \geq 1)$	(4) $\Pr(K_i \geq 2)$
Privacy concern (1–7)	−0.008 (0.006)	+0.000 (0.001)	−0.007 (0.006)	+0.000 (0.001)
Digital literacy (sum of 5 items)	−0.005 (0.006)	−0.001 (0.001)	−0.006 (0.006)	−0.001 (0.001)
Log net monthly income	+0.061*** (0.006)	+0.008*** (0.001)	+0.070*** (0.007)	+0.010*** (0.001)
Education (CBS 1–6)	−0.015 (0.010)	−0.002* (0.001)	−0.014 (0.010)	−0.002 (0.001)
Tertiary educ. (≥5)	+0.054*** (0.018)	+0.007*** (0.003)	+0.049*** (0.019)	+0.007** (0.003)
Age (years)	+0.071** (0.032)	+0.013*** (0.004)	+0.095*** (0.036)	+0.018*** (0.005)
Age^2	−0.080** (0.032)	−0.013*** (0.005)	−0.101*** (0.035)	−0.018*** (0.005)
Mean of outcome	0.273	0.038	0.273	0.038
N	3,116	3,116	3,116	3,116
R^2/pseudo-R^2	0.049	0.046	0.023	0.009

Notes: The dependent variables are the cumulative threshold probabilities $\Pr(K_i \geq 1) = \hat{p}_{i,1} + \hat{p}_{i,2}$ and $\Pr(K_i \geq 2) = \hat{p}_{i,2}$. Columns (1)–(2): linear-probability OLS with HC1-robust standard errors in parentheses. Columns (3)–(4): average marginal effects from a fractional (Bernoulli quasi-likelihood) logit on the same outcomes and regressors. Continuous regressors are z-scored, so coefficients are per within-sample standard deviation. All columns use the same maximal sample ($N = 3,116$). $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

E Welfare Decomposition Identities

Decision loss. Let $p = \Lambda(V^*)$ be the sharing probability of a consumer with correct beliefs. For any sharing probability q,

$$q\,V^* + H(q) = \log\bigl(1 + e^{V^*}\bigr) - \mathrm{KL}(q \,\|\, p), \qquad \mathrm{KL}(q \,\|\, p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \;\geq\; 0,$$

which follows from $\log p = V^* - \log(1 + e^{V^*})$ and $\log(1-p) = -\log(1 + e^{V^*})$. Welfare in equation (19) is therefore the logit surplus with correct beliefs minus a decision loss, the Kullback–Leibler divergence between the consumer's sharing probability and the one she would choose with correct beliefs. The loss is zero only when $q = p$.

Decision channel. Holding the baseline realized values $V^{*,0}$ fixed, the decision channel in equation (23) equals

$$U_i(q^{\mathrm{cf}}, V^{*,0}) - U_i(q^0, V^{*,0}) = \sum_s \Bigl[\mathrm{KL}\bigl(q^0_{is} \,\|\, p^0_{is}\bigr) - \mathrm{KL}\bigl(q^{\mathrm{cf}}_{is} \,\|\, p^0_{is}\bigr)\Bigr],$$

the reduction in the decision loss relative to the baseline market, with $p^0_{is} = \Lambda(V^{*,0}_{is})$. It is positive when the counterfactual choices are closer to the best response to the baseline market, and can be negative for consumers whose forecasts move away from it.

Equilibrium channel. Because η^*_{is} is linear in g_i,

$$\eta^{*,\mathrm{cf}}_{is} - \eta^{*,0}_{is} = -\psi_s\bigl[(g_i - \bar{g}^{*,\mathrm{cf}}_s) - (g_i - \bar{g}^{*,0}_s)\bigr] = \psi_s\bigl(\bar{g}^{*,\mathrm{cf}}_s - \bar{g}^{*,0}_s\bigr),$$

so the equilibrium channel equals $\gamma_1 \sum_s q^{\mathrm{cf}}_{is}\, \psi_s(\bar{g}^{*,\mathrm{cf}}_s - \bar{g}^{*,0}_s)$ and depends on the consumer only through her sharing probabilities.

Welfare relevance. The welfare-relevant value in equation (25) satisfies $V^*_{is}(\lambda) = (1-\lambda)V^{*,\mathrm{cf}}_{is} + \lambda V^{*,0}_{is}$. Because choices do not depend on λ and welfare is linear in V^* given choices, $\Delta U_i(\lambda)$ equals the decision channel plus $(1-\lambda)$ times the equilibrium channel, and

the aggregate effect changes sign at $\lambda^* = 1 + \overline{\text{decision}}/\overline{\text{equilibrium}}$.

F Additional Counterfactual Results

Table 17. Bootstrap confidence intervals for the counterfactual statistics

	Estimate	Bootstrap s.e.	95% interval
Increase in sophistication, $\hat{\tau} \to \tau_{\text{cf}} = 3$			
Share naive, π_0	0.728	0.016	[0.703, 0.763]
Change in mean welfare, ΔW	-0.087	0.021	[-0.132, -0.057]
in percent of baseline welfare	-4.3	1.3	[-7.1, -2.7]
Decision channel	+0.092	0.024	[+0.061, +0.140]
Equilibrium channel	-0.179	0.044	[-0.272, -0.118]
Respondents who gain (%)	11.1	0.8	[10.0, 13.0]
Welfare-relevance threshold λ^*	0.487	0.008	[0.466, 0.495]
RE welfare minus baseline welfare	-0.102	0.024	[-0.152, -0.067]
Naive subsidy, $D_s = \bar{g}_s^*(\hat{\tau}) - \bar{g}_s^{\text{RE}}$			
Statistics bureau	-0.273	0.037	[-0.347, -0.218]
Researchers	-0.296	0.040	[-0.369, -0.228]
Government	-0.250	0.036	[-0.323, -0.207]
Internet provider	+0.067	0.028	[+0.020, +0.118]
Online shops	+0.272	0.050	[+0.176, +0.373]
Market research	-0.213	0.036	[-0.298, -0.152]
App companies	+0.214	0.044	[+0.141, +0.310]
Tech companies	+0.441	0.099	[+0.303, +0.663]
Social media	+0.459	0.100	[+0.319, +0.646]

Notes. Respondent-clustered bootstrap with 50 draws of the structural parameters. At each draw, the posteriors over cognitive levels and privacy costs and all counterfactual statistics are recomputed on the estimation sample. 95 percent percentile intervals. Welfare in log-odds units of the sharing decision.

Table 18. Counterfactual under the model with covariate-dependent sophistication

	Main model	Covariate model
Share naive, baseline	0.728	0.738
Share naive, counterfactual	0.118	0.118
Change in mean welfare, ΔW	-0.087	-0.069
in percent of baseline	-4.3	-3.4
Decision channel	+0.092	+0.072
Equilibrium channel	-0.179	-0.142
Respondents who gain (%)	11.1	13.7
bottom income decile	88.2	91.1
top income decile	22.3	42.7
Welfare-relevance threshold λ^*	0.487	0.490
RE welfare minus baseline	-0.102	-0.099
Naive subsidy, technology companies	+0.441	+0.421
Naive subsidy, researchers	-0.296	-0.295

Notes. Main model: common cognitive intensity (Section 8). Covariate model: $\log \tau_i = \delta_0 + X_i'\delta$ (Section 7.1); the policy raises δ_0 so that the share of naive consumers equals that of the main counterfactual. Welfare in log-odds units of the sharing decision.

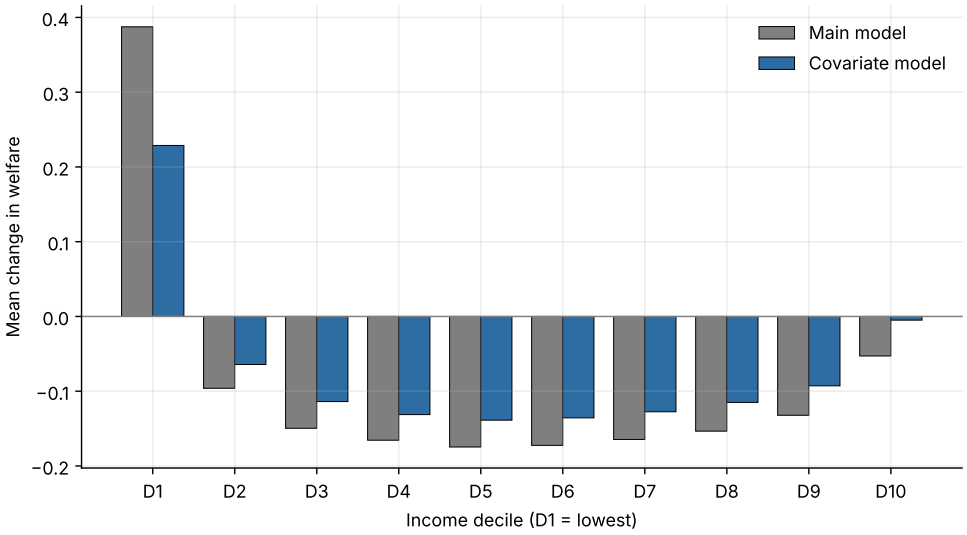


Figure 15. Welfare change under increasing sophistication by income decile: main model and model with covariate-dependent sophistication.

Table 19. Welfare relevance of sorting, by organization

Organization	$\hat{\psi}_s$	Decision	Equilibrium	ΔW $(\lambda = 0)$	λ_s^*
Statistics bureau	-3.06	+0.017	-0.030	-0.014	0.45
Researchers	-3.40	+0.020	-0.036	-0.016	0.44
Government	-2.68	+0.013	-0.023	-0.010	0.45
Internet provider	+0.54	+0.001	-0.001	-0.001	0.51
Online shops	+1.90	+0.007	-0.014	-0.008	0.53
Market research	-2.04	+0.006	-0.011	-0.005	0.46
App companies	+1.48	+0.004	-0.008	-0.004	0.52
Tech companies	+2.78	+0.013	-0.029	-0.016	0.54
Social media	+2.76	+0.012	-0.027	-0.014	0.54
All		+0.092	-0.179	-0.087	0.49

Notes. Contributions of each organization to the mean change in welfare per respondent; contributions sum to the totals in Table 10. λ_s^* is the share of the type slope that reflects real surplus at which the organization's contribution changes sign.

Table 20. Sharer-pool averages along the sophistication path, by organization

Organization	$\hat{\psi}_s$	$\bar{g}_s^*(0)$	$\bar{g}_s^*(\hat{\tau})$	$\bar{g}_s^*(\tau_{\text{cf}})$	$\bar{g}_s^{\text{RE}}$	D_s
Statistics bureau	-3.06	-0.004	+0.052	+0.273	+0.326	-0.273
Researchers	-3.40	-0.004	+0.059	+0.297	+0.356	-0.296
Government	-2.68	-0.004	+0.047	+0.249	+0.296	-0.250
Internet provider	+0.54	-0.005	-0.015	-0.071	-0.082	+0.067
Online shops	+1.90	-0.005	-0.042	-0.269	-0.314	+0.272
Market research	-2.04	-0.005	+0.038	+0.213	+0.251	-0.213
App companies	+1.48	-0.005	-0.034	-0.214	-0.248	+0.214
Tech companies	+2.78	-0.005	-0.059	-0.418	-0.500	+0.441
Social media	+2.76	-0.005	-0.060	-0.436	-0.520	+0.459

Notes. Realized pool averages $\bar{g}_s^*$ when all consumers are naive ($\tau \to 0$), at the estimate $\hat{\tau}$, at $\tau_{\text{cf}} = 3$, and under rational expectations. D_s is the naive subsidy.