Designing Choice Architecture to Mitigate Selection Bias in Consumer Data Sharing

TESARY LIN
Boston University
and
AVNER STRULOV-SHLAIN
University of Chicago

Choice architecture is widely used to nudge consumers into sharing data in consent-based data exchanges. We present experiment evidence from Lin and Strulov-Shlain [2023] demonstrating that conventional choice architecture design could lead to biases in sample data. We illustrate how the tension between maximizing data volume and minimizing data bias depends on both supply and demand factors. We also highlight the need for organizations to consider both the volume and representativeness of sample data when optimizing their choice architecture for data collection.

Categories and Subject Descriptors: [**Applied computing**]: Law, social and behavioral sciences—*Economics*; [**Security and privacy**]: Human and societal aspects of security and privacy—*Economics of security and privacy*

General Terms: Design, Economics, Experimentation, Human factors, Management, Measurement

Additional Key Words and Phrases: Privacy, Choice architecture, Market for data, Selection bias

1. INTRODUCTION

Empirical science and business analytics often encounter scenarios where the data used for analysis are collected in a way that deviates from a random sampling procedure. Such data cause a distorted representation of the underlying population that the analytics intends to learn about. The resulting bias in the estimates and inference, called selection bias, is prevalent when subjects can choose whether to share their data.

Unrepresentative data can severely degrade the quality of insights and subsequent decision-making. In clinical trials, for instance, the lack of ability to recruit minority participants leads to noisy estimates of new treatments' efficacy and side effects that these minorities experience. Mayo Clinic reports that these imprecise estimates led to increases in US healthcare expenditure by $1.2 trillion in 2003-2006 [Ma et al. 2021]. Businesses face similar problems when their customers differ in product preferences while only a selected subsample gives feedback [Blattner and Nelson 2021; Cao et al. 2021].

Nevertheless, current managerial strategies to encourage consumer data sharing primarily focus on ensuring sufficient volume. These strategies often involve "consent engineering" practices, also known as nudges, choice architecture, or dark patterns, to maximize customer consent rates [Utz et al. 2019; Nouwens et al.

Authors' addresses: tesary@bu.edu, avner.strulov-shlain@chicagobooth.edu

2020]. As one of the biggest consent management platforms, OneTrust promoted its "consent rate optimization" product in a company blog article:[1]

> Take advantage of Consent Rate Optimization to **maximize consent rates** through advanced A/B testing,...

However, these practices show little concern about the composition of consented customers and whether they are representative of the underlying customers.

In this letter, we present findings from Lin and Strulov-Shlain [2023], which examines how corporate nudging practices in data exchange settings and their focus on data volume affect selection bias. We characterize the trade-off between volume maximization and bias mitigation, and describe how the trade-off depends on both the demand and supply of consumer data. We use "data markets" to refer to settings where organizations offer products or prices in exchange for consumers' consent for data collection and processing.

2. WHEN DOES SELECTION BIAS HINDER DATA-DRIVEN DECISION QUALITY?

The value of data lies in its ability to provide valuable information for decision making. Therefore, data users should care about selection bias when it negatively affects the quality of predictions and inferences derived from these data. A biased dataset can compromise the accuracy of data-driven insights in the following scenarios:

(1) The data user wants to learn the average outcome, but the sample data is unrepresentative, and the data user does not have enough information to reweigh the sample data to recover an unbiased estimate. As an example, researchers at the Urban Institute recently surveyed economists to gather their attitudes about privacy protection procedures applied to census data [Williams et al. 2024]. Only 4% of economists returned their survey. Furthermore, while the researchers have the demographics of the respondents who returned the survey, they do not know the demographic distribution across the economic profession, so they cannot use demographics to reweigh the survey responses. Sample reweighting is not a panacea: It decreases the effective sample size and increases estimates' variance. This problem is exacerbated as the variance of weights increases [Stantcheva 2022], that is, as the sample data become more biased compared to the population of interest.

(2) The data user wants to learn the heterogeneity of the outcome, but the sample data under-represent certain subgroups, preventing effective learning about outcomes from this subgroup. The clinical trial example aptly illustrates this point: Even though medical researchers know how the participants in their trial differ from the general population in its demographic distribution, they cannot use this information to mitigate the fact that drug effects on under-represented subgroups are estimated with lower precision.

It is natural to ask if there are exceptions where unrepresentative data still yield unbiased outcome estimates. In theory, this scenario is likely when consumers have

[1] https://www.onetrust.com/blog/onetrust-launches-consent-rate-optimization-to-maximize-opt-ins/

similar preferences and behaviors that the data user wants to predict. In practice, this is rarely true in many applications.

3. EXPERIMENT: PRIVACY VALUATIONS ACROSS SUB-POPULATIONS UNDER THE INFLUENCE OF CHOICE ARCHITECTURE

To understand how companies' choice architecture design affects consumers' data sharing decisions and the resulting quality of sample data, we need to observe variations in choice architecture that do not correlate with other factors that can contribute to differences in data sharing decisions. We also need to observe customer characteristics to understand how different consumer subgroups respond to the same choice architecture differently.

To achieve these goals, we design an experiment that randomizes participants (recruited via Facebook Ads and Prolific) into different choice architecture designs while we elicit their valuations for their private Facebook data. We use multiple price lists to elicit data values [Andersen et al. 2006], which is a common tool for measuring incentive-compatible valuations when the products under consideration do not have a known market price, such as free digital goods [Brynjolfsson et al. 2019]. One well-known fact about privacy valuations is their context specificity: Consumers can show varied preferences based on who has access to their data and how they will be used [Martin and Nissenbaum 2016; Lin 2022]. To separate the variation induced by different economic and social contexts from variations induced by nudges, we specifically ask participants how much compensation they are willing to accept for sharing data with advertisers.

To introduce variations in choice architecture, we include different default settings and price anchors on the multiple price list interface as our treatment. We then search for the design combination that maximizes data volume (hereafter, the volume-maximizing design) and the one that minimizes data bias (the bias-minimizing design) based on participant responses. To collect consumer characteristics, we ask participants the demographic groups they belong to, as well as their web and social media consumption habits.

In the spirit of Ludwig et al. [2011], we view our artefactual experiment as a mechanism probe, rather than a policy evaluation project. The experiment allows us to see how the trade-off between volume maximization and bias mitigation depends on both the demand and supply of consumer data, which we explain below.

4. MAIN FINDINGS

4.1 Privacy valuations vary, yet choice architecture has a substantial influence

Figure 1 shows the data supply curves across our six choice architecture design treatments. Consumers vary substantially in how much they value their own Facebook data: The average valuation is $67, but certain participants indicate a willingness-to-accept in the magnitude of thousands, and 18% of them indicate their unwillingness to share data with advertisers at any price.

Despite the wide range of privacy values, choice architecture substantially influences the data value distribution. The opt-out design decreases consumers' privacy valuations by 13.6% on average compared to an opt-in frame. The effect of price anchor is even larger: Switching from a $50-100 price anchor to a $0-50 anchor

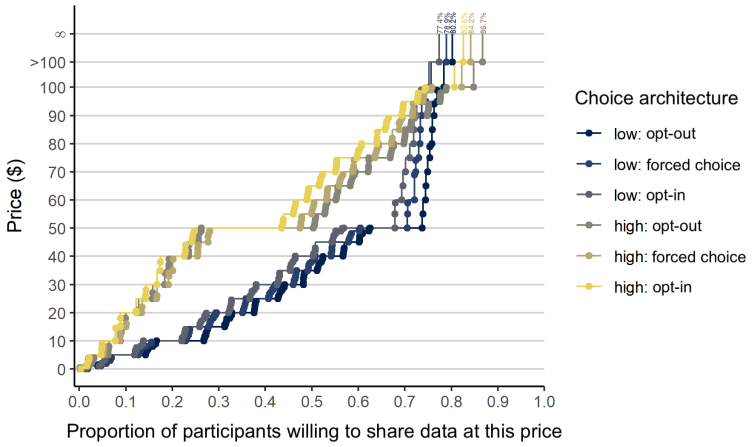


Fig. 1: Data supply curves across choice architecture treatments

decreases data valuation by 52.6% on average.

4.2 Negative correlation between privacy valuation and choice architecture effects in certain domains

A key supply-side contributor to the tension between volume-maximizing and bias-minimizing goals is the negative correlation between consumers' initial privacy valuations and their responsiveness to nudges. Imagine two groups of consumers: high-income and low-income. Low-income consumers, while valuing their privacy less in the absence of choice architecture, are more susceptible to its influences compared to their wealthier counterparts. When not deploying choice architecture, companies would have undersampled wealthier customers while oversampling poorer ones. With a volume-maximizing choice architecture that encourages more data sharing, selection bias may intensify as low-income consumers are disproportionately influenced.

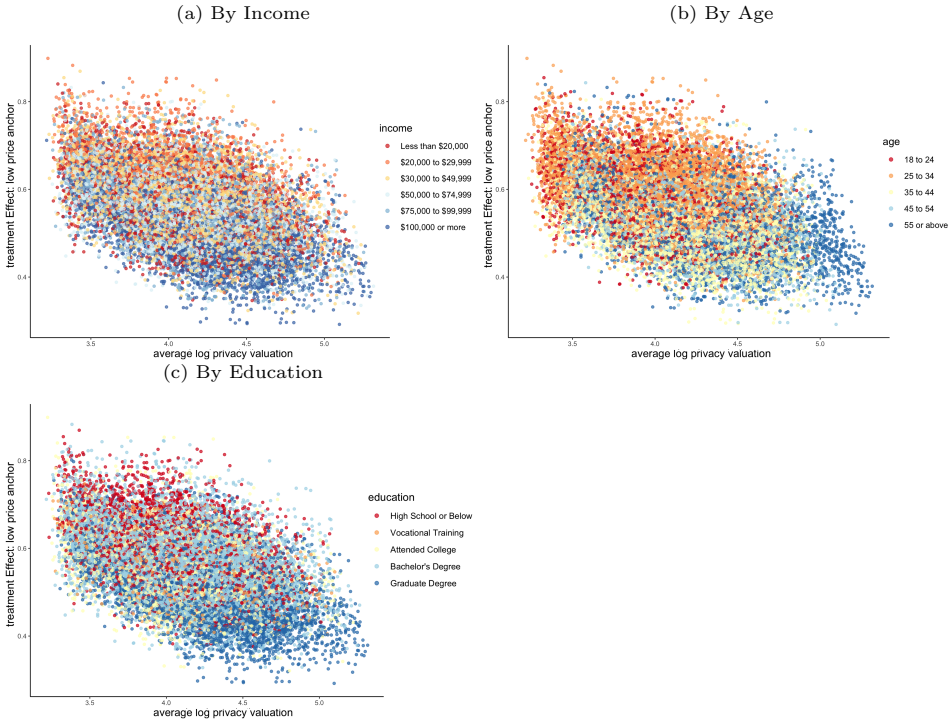
To illustrate this point empirically, we focus on demographic subgroups where such a negative correlation exists between privacy valuation and choice architecture effects. In our setting, younger, poorer, and less educated consumers tend to value their personal data less, but are also more responsive to the influence of choice architecture, as shown in Figure 2.

4.3 The volume-bias trade-off: supply and demand factors

Although the joint distribution of privacy valuation and choice architecture is a key supply-side contributor to the volume-bias trade-off, the manifestation of this trade-off also depends on the demand side. Here, we focus on two demand-side factors: (a) the elasticity of demand; (b) the current data price relative to the privacy value distribution.

Inelastic demand exacerbates the volume-bias trade-off. A volume-maximizing choice architecture shifts the data supply curve outwards, meaning more consumers are sharing data at each price point. With inelastic demand, the firm adjusts its price for data downwards in response to the outward shift in supply, rather than

Fig. 2: Negative correlation between privacy valuation and response to nudges



maintaining the current price to draw in more consumers. This scenario is possible when firms perceive the marginal value of data to decline fast. With an inelastic demand, we are more likely to see biased datasets as a result of deploying the volume-maximizing choice architecture.

Conversely, a firm with elastic demand keeps its price for data similar to the level without choice architecture, while gathering more data from consumers. Sometimes firms can mitigate selection bias simply by gathering more data. As an extreme example, suppose initially all low-income consumers already share their data with the firm while all high-income people remain unwilling to share data. The volume-maximizing choice architecture, by encouraging data sharing among both low and high-income consumers at the initial price point, reduces bias as high-income consumers are now more likely to included in the dataset.

Higher data prices facilitate the alignment between volume-maximizing and bias-mitigating goals. The example above shows that when the over-sampled group has already fully opted in, a volume-maximizing design can indirectly mitigate selection bias by drawing in more consumers from the under-sampled group. This scenario is more likely to hold when the price for data is high compared to the privacy value distribution. In other words, for companies that compensate consumers well for their data, the volume-maximizing and bias-mitigating designs are more likely to coincide.

Figure 3 illustrates how the volume and bias comparison across choice archi-

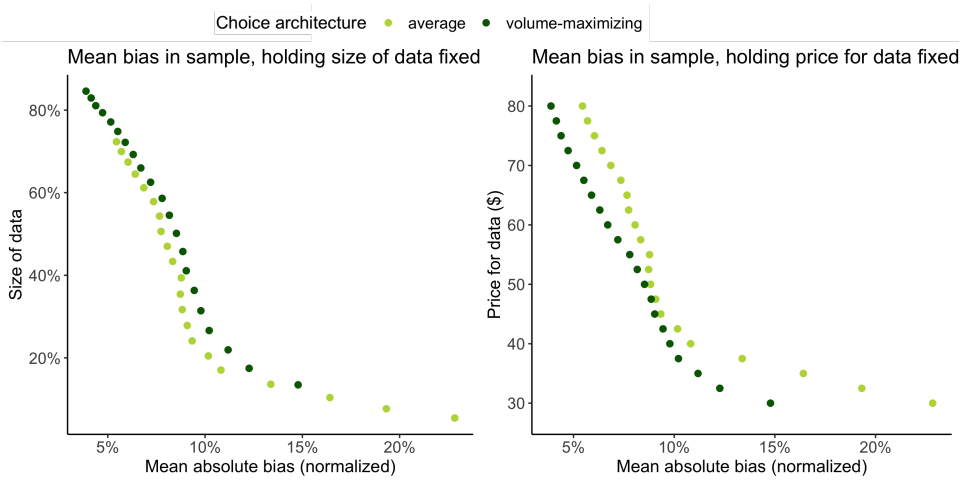


Fig. 3: Data supply curves across choice architecture treatments

tecture designs depend on both demand elasticity and data market prices. We estimate counterfactual data-sharing decisions under different choice architecture designs and data prices using causal forests, then compare data quality under a volume-maximizing and a benchmark "randomly" choice architecture. In the left panel, we align pairs of sample data based on their volume, representing the case where companies' demand for data is perfectly inelastic and thus the equilibrium quantity of data traded is the same. On the right panel, we connect pairs of sample data collected under the same price, representing the other extreme where companies' demand for data is perfectly inelastic and thus the equilibrium price for data is constant. Consistent with our previous argument, the volume-maximizing design exacerbates sample bias (in terms of demographics) in the inelastic demand condition, and mitigates sample bias in the elastic demand setting.

4.4 The role of choice architecture personalization

Personalization is a common feature when companies optimize their choice architecture. In our opening example, OneTrust provides not only A/B testing services to measure the performance of a design, but also targeting capabilities based on attributes "such as behavior, age, content and more." On the other hand, personalizing prices or equivalent offers in data-sharing settings is often prohibited by existing privacy regulations such as GDPR and CCPA. It is natural to ask how the presence of design personalization changes the bias-volume trade-off when firms need to provide uniform pricing.

In Lin and Strulov-Shlain [2023], we show that choice architecture personalization increases its efficacy in reducing selection bias, holding the volume target fixed. In our setting, personalized designs are more effective in in reducing bias because they give the company more granular control in deciding the mix of consumers drawn in at any given price point. In comparison, personalization leads to limited gains in volume maximization, because the design combinations that increase data sharing the most are often (though not always) the same across participants. In

combination, the ability to use personalized design can lead the company to favor bias reduction over volume maximization as the former becomes relatively more effective.

5. CONCLUSION

Choice architecture is prevalent in consent-based data exchange markets. We show that conventional choice architecture optimization practices focusing solely on maximizing data volume can negatively impact data quality by exacerbating sample selection bias. We argue that companies and organizations should consider the bias-volume trade-off when designing and deploying choice architecture to improve the quality of data collected for decision making.

REFERENCES

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. 2006. Elicitation using multiple price list formats. *Experimental Economics 9,* 4, 383–405.

Blattner, L. and Nelson, S. 2021. How costly is noise? data and disparities in consumer credit. *arXiv preprint arXiv:2105.07554*.

Brynjolfsson, E., Collis, A., and Eggers, F. 2019. Using massive online choice experiments to measure changes in well-being. *Proceedings of the National Academy of Sciences 116,* 15, 7250–7255.

Cao, R., Koning, R. M., and Nanda, R. 2021. Biased sampling of early users and the direction of startup innovation. *NBER Working Paper No. 28882*.

Lin, T. 2022. Valuing intrinsic and instrumental preferences for privacy. *Marketing Science*.

Lin, T. and Strulov-Shlain, A. 2023. Choice architecture, privacy valuations, and selection bias in consumer data. In *Proceedings of the 24th ACM Conference on Economics and Computation*. 960–960.

Ludwig, J., Kling, J. R., and Mullainathan, S. 2011. Mechanism experiments and policy evaluations. *Journal of economic Perspectives 25,* 3, 17–38.

Ma, M. A., Gutiérrez, D. E., Frausto, J. M., and Al-Delaimy, W. K. 2021. Minority representation in clinical trials in the united states: trends over the past 25 years. In *Mayo Clinic Proceedings*. Vol. 96. Elsevier, 264–266.

Martin, K. and Nissenbaum, H. 2016. Measuring privacy: an empirical test using context to expose confounding variables. *Colum. Sci. & Tech. L. Rev. 18*, 176.

Nouwens, M., Liccardi, I., Veale, M., Karger, D., and Kagal, L. 2020. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

Stantcheva, S. 2022. How to run surveys: A guide to creating your own identifying variation and revealing the invisible. *Annual Review of Economics 15*.

Utz, C., Degeling, M., Fahl, S., Schaub, F., and Holz, T. 2019. (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*. 973–990.

Williams, A., Snoke, J., Bowen, C., and Barrientos, A. 2024. Disclosing economists' privacy perspectives: A survey of american economic association members' views on differential privacy and the usability of noise-infused data. *Harvard Data Science Review*.