The Economic Engineering of Personalized Experiences

by

Andreas A. Haupt

BS in Mathematics, University of Bonn, 2014
MS in Mathematics, University of Bonn, 2017
MS in Economics, University of Bonn, 2018
BS in Computer Science, University of Frankfurt, 2019

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ENGINEERING-ECONOMIC SYSTEMS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored by:	Andreas A. Haupt Institute for Data, Systems, and Society January 31, 2025
Certified by:	Alessandro Bonatti Professor of Applied Economics, MIT
Certified by:	Dylan Hadfield-Menell Professor of Computer Science, MIT
Certified by:	Eric Maskin Professor of Economics and Mathematics, Harvard University
Certified by:	David Parkes Professor of Computer Science, Harvard University
Accepted by:	Fotini Christia Professor of the Social Sciences Graduate Officer, Institute for Data, Systems, and Society

The Economic Engineering of Personalized Experiences

by

Andreas A. Haupt

Submitted to the Institute for Data, Systems, and Society
on January 31, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ENGINEERING-ECONOMIC SYSTEMS

ABSTRACT

Consumer applications employ algorithms to deliver personalized experiences to users, among others, in search, e-commerce, online streaming, and social media, impacting how users spend their time and money. This dissertation studies the design of such personalization algorithms and the economic consequences of their deployment.

The first chapter focuses on the impacts of reward signal precision on online learning algorithms frequently used for personalization. Reward signals are precise when individual measurement is accurate and heterogeneity is low. While some algorithms, which we call "risk-averse", favor experiences that yield more precise reward signals and hence favor measurability and homogeneity, others, in the limit, choose experiences independently of the precision of their associated reward signals.

The third chapter analyzes how preference measurement error differentially affects user groups in optimal personalization. If such measurement error is symmetric, welfare maximization requires delivering majority-preferred experiences at a rate beyond their proportion in the user population and hence increasing concentration. However, asymmetric preference measurement errors may arise due to users' actions to reduce measurement error. Participants in a survey of TikTok state that they engage in such costly actions.

The fifth chapter studies, through the introduction of a new desideratum for market design, how to achieve personalization without infringing on user privacy. Contextual privacy demands that all (preference) information elicited by an algorithm is necessary for computing an outcome of interest in all possible configurations of users' information. This property is demanding, as it requires that no two pieces of information can jointly but not unilaterally influence the outcome. Algorithms can protect the privacy of users who are queried late and whose information is not used to compute public statistics of the user population, hence achieving the relaxed notion of maximal contextual privacy.

Two brief chapters introduce new models of human-machine interaction. The first examines the design of generative models, while the second proposes stated regret of past consumption as a new data modality and presents a corresponding data collection tool.

Alessandro Bonatti
Professor of Applied Economics, MIT

Dylan Hadfield-Menell
Professor of Computer Science, MIT

Eric Maskin
Professor of Economics and Mathematics, Harvard University

David Parkes
Professor of Computer Science, Harvard University

Acknowledgments

Completing this dissertation has been a challenging yet deeply rewarding journey, and I am grateful to many individuals whose support made this possible.

I am profoundly appreciative of the guidance and mentorship of my Prof. Alessandro Bonatti and Prof. Dylan Hadfield-Menell throughout my PhD. Both taught me through my first papers in Economics and Artificial Intelligence, respectively, and helped me understand what it means to work in these fields. I also thank my committee members, Prof. Eric Maskin and Prof. David Parkes, for their feedback and mentorship over the past five years. Additionally, I wish to thank my teachers and mentors who supported me along this journey, including Prof. Mohammad Akbarpour, Prof. Erik Brynjolfsson, Prof. Ashton Carter (late), Prof. Benjamin Golub, Bernhard Kötter, Prof. Shengwu Li, Prof. Mathias Risse, Prof. Parag Pathak, and Dr. Ngoc Mai Tran.

I would like to acknowledge the support I received through the MIT Presidential Fellowship, the Dahleh Group, and the Algorithmic Alignment Group, which enabled me to pursue this research. I am also grateful to the MIT Department of Economics and the MIT Sloan School of Management for the opportunities to present my research in Industrial Organization and Microeconomic Theory in lunch seminars. I also thank the former Unit C.6 for Data Markets and E-Commerce at the European Commission's Directorate-General for Competition and the Office of International Affairs at the Federal Trade Commission for invaluable policy-related experiences. I am grateful to Elizabeth Milnes and Teresa Coates-Cataldo for their excellent administrative support. Special thanks to my colleagues and professional supervisors, including Philip Christoffersen, Maria Coppola, Andrew J. Koh, Jenny Romelsjö, Paul O'Brien, Mark York, and Stewart Slocum.

I am thankful to my collaborators Dr. Mihaela Curmei, Prof. Dylan Hadfield-Menell, Dr. Zoë Hitzig, Aroon Narayanan, and Prof. Chara Podimata for working with me on articles partly or wholly included in this dissertation. Their insights and collaboration have greatly enriched my understanding of game theory, online learning, and software development. I am also grateful to my collaborators on additional research projects during my PhD, including Prof. Dirk Bergemann and Prof. Alex Smolin (Bergemann, Bonatti, A. Haupt, et al. 2021), Dr. Xiaotong Guo and Prof. Hai Wang (X. Guo et al. 2023), Prof. Gabriele Farina and Brian Zhang (Zhang, Farina, Anagnostides, Cacciamani, S. M. McAleer, et al. 2023; Zhang, Farina, Anagnostides, Cacciamani, S. McAleer, et al. 2024), Dr. Nicole Immorlica and Dr. Brendan Lucier (A. A. Haupt, Immorlica, and Lucier 2024), Dr. Ian Gemp (Gemp et al. 2025) and Olivia Hartzell (Hartzell and A. Haupt 2025). Many of them have become friends and continue to inspire me.

Beyond academia, I am deeply grateful to my friends and family for their unwavering support. To my mother, Helga Haupt, my sister, Katrin Haupt, and my godparents, thank you for encouraging and supporting me as I pursued a path far from the one you followed. I would also like to acknowledge my late father, Dr. Klaus Schmidt, who saw great potential in me yet was on this planet too briefly to witness its realization.

Finally, I am fortunate to have been surrounded by many wonderful friends whose love, care, and support have sustained me through these years: Samuel Bosch, Marieke de Bock, Bernardo Garcia Bulle Bueno, Alexander Haberman, Julie Heng, Rachael Kha, Eli Kramer, Sirui Li, Marieke Meier, Yanika Meyer-Oldenburg, Hannah Morin, Amir Nouripour, John-Alexander Preuß, Eva Richter, Juan-Pablo Rodriguez, Charlotte Siegmann, Olivia Siegel, Max Simchowitz, Stewart Slocum, Anne Snyder, Stephen Zoepf, Mark York, and Taimor Williams.

To all who have been part of this journey—thank you.

Biographical Sketch

Andreas Haupt is a PhD Candidate in Engineering-Economic Systems at the Institute for Data, Systems, and Society (IDSS) and the Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT). Originally from Frankfurt, Germany, he studied Mathematics, Economics, and Computer Science in Bonn, Lausanne, and Frankfurt. Andreas joined MIT in the Fall of 2019 and has researched digital platforms, preference elicitation, and mechanism design. He was a trainee at the European Commission's Directorate-General for Competition in the spring of 2021 and an intern at the United States Federal Trade Commission in the summer of 2023. His work has been presented at the ACM Conference on Economics and Computation and published in *Games and Economic Behavior*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning systems increasingly shape user behavior by tailoring experiences to diverse audiences. The emergence of large-scale datasets has enabled algorithms to personalize experiences at an unprecedented scale, influencing billions of users. From search engines and e-commerce platforms to online streaming and social media, personalization algorithms play a central role in determining how users interact with content, make purchasing decisions, and allocate their time.

Personalization algorithms do more than tailor content to individual users—they shape markets by influencing both the demand and supply of goods. When a recommendation system consistently favors certain products or content over others, it can alter user behavior, influence market structure, and amplify market concentration. While the effect of a single choice of a personalized experience may seem negligible, its impact scales dramatically when applied to billions of users, effectively steering aggregate demand.

These market-wide consequences mean that engineers designing personalization algorithms must consider consider their economic impacts. The ability of algorithms to shape the demand and to favor concentration or diversity necessitates engineering that has market consequences in mind. This thesis engages in such *Economic Engineering of Personalized Experiences*.

The thesis makes three main contributions. First, we analyze the concept of "algorithmic demand", examining its dependence on noise in preference measurement and the choice of learning algorithms (Chapters 2 and 4). Second, we formalize privacy in algorithmic personalization through the lens of *contextual integrity* (Nissenbaum 2004; Benthall, Gürses, and Nissenbaum 2017) and establish its relationship to other desiderata (Chapter 6). Finally, we study two alternative modes of user interaction with algorithms. First, we propose data on *regret of previous consumption* as a new data modality of retrospective evaluations (Chapter 5). Second, we examine the design of *generative models* and their personalization through users drawing a different number of samples (Chapter 3).

1.1 The Intersection of Two Fields

Personalization has been extensively studied in both Economics and Artificial Intelligence, with different perspectives on the problem. In Economics, Industrial Organization and

Mechanism Design focus on the market and distributional consequences of algorithms, and their incentive properties, but abstract from sequential decision-making or imperfect and incomplete preference revelation. Conversely, Artificial Intelligence and decision-making provide strong theoretical guarantees for learning and optimization and empirically performant algorithms, but typically abstracts from broader economic and market effects. While both fields aim to design technologies that facilitate human economic activity, they grapple with the challenge of developing tractable models that capture the complexity of the market effects of personalization algorithms.

Bridging mechanism design and online learning is essential for effective personalization. The field of Artificial Intelligence provides practical algorithms and accounts for technological constraints, while Economics offers modeling tools to analyze the broader market effects of these algorithms. Integrating these perspectives enables the design of personalization systems that are both computationally efficient and aware of their economic impacts.

1.2 Personalized Introductions

To make this thesis accessible to readers with backgrounds in Artificial Intelligence or Economics, I provide separate, personalized introductions for each field. Readers may choose to engage with the introduction most relevant to their expertise or read both for a more comprehensive perspective. A discussion of related literature is included in individual chapters to maintain focus on the core arguments in this introduction.

1.2.1 A Personalized Introduction for Artificial Intelligence Researchers

Several parts of this thesis can be understood as decision-making under uncertainty, where the uncertainty is a combination of *heterogeneity* in preferences and *preference measurement error*. Chapters 2 and 4 fall into the framework of a *correlated stochastic bandit with a latent random source* (S. Gupta, Joshi, and Yağan 2020), where the correlation structure is governed by a latent user type θ. The problem is defined by:

- Θ: a latent type space representing user preferences
- $\theta \sim F \in \Delta(\Theta)$: a prior distribution over user types
- A: an action space containing personalized experiences
- $R\colon A \times \Theta \to \mathbb{R}_+$: a stochastic reward function, where rewards for different actions are correlated via θ
- $O\colon A \times \Theta \to \Omega$: an observation model, providing noisy feedback about the reward

At each round t, a user of type θ_t is being served, the algorithm selects a personalized experience $a_t \in A$, and observes a noisy reward $O(a_t, \theta_t)$. The goal is to learn an optimal *policy* $\pi\colon O \to \Delta(A)$ that maximizes the expected cumulative discounted reward $\mathbb{E}\left[\sum_{t=0}^{T} R(a_t, \theta)\right]$. To conform with conventions in mechanism design, we will sometimes write $x \in X$ instead of $a \in A$ to denote the algorithm's decision, and write $u(x;\theta)$ instead of $R(x,\theta)$ for the reward. $O(x;\theta)$ is a noisy measurement of the utility, *e.g.*,

$$O(x;\theta) = R(x;\theta) + \varepsilon_t(x,\theta).$$

We call ε_t the *preference measurement error*. From the perspective of the algorithm, observations will be random conditional on a realization of ε, based on the variation in θ_t.

Chapters 2 and 4 analyze the properties of heuristic and optimal policies π in the context of choosing personalized experiences.

Chapter 2 demonstrates that engineering choices, even those that seem inconsequential under asymptotic guarantees, can significantly influence the probabilities of selecting different actions x. We refer to the probability distribution on actions $x \in X$ as the *algorithmic demand* of policy π, since in many settings, the algorithm directly (by displaying content to users) or indirectly (by recommending products that may lead to purchases) shapes realized demand.

Chapter 4 further investigates algorithmic demand, now for the optimal algorithm with knowledge of the exact probability generating model, and varying the preference measurement error ε_t. Under arbitrary noise, we find that, as in Chapter 3, both over- and under-representation of statistical minorities can occur. Under a notion of symmetry we make precise, however, personalization increases concentration.

The brief Chapter 3 explores a setting where user-algorithm interactions are severely constrained. Inspired by generative models, it considers a case in which in each episode, the algorithm makes no observations, but a user can continue to draw samples from a designed distribution, at a cost $c > 0$. In this model, all personalization is driven by user decisions. While in a one-shot interaction, a deterministic solution is optimal for the designer, we show that randomization and over-representation of minorities are possible.

The last two chapters take a broader perspective on the interaction between algorithms and users, exploring how personalization systems should be designed to align with user preferences and privacy considerations.

The brief Chapter 5 examines a classical question from alignment (Hadfield-Menell 2021): does the reward function R truly capture desirable outcomes for users? It argues that utility $R(s,a)$ should be re-evaluated after an interaction with a personalized experience. To facilitate this, the chapter introduces software for collecting user feedback on regrets arising from past recommendations, with a case study focusing on the personalized experiences on youtube.com.

The final chapter, Chapter 6, investigates a fundamental requirement for *private* personalization: any observation made by the policy should be necessary to achieve an intended goal. We show that fully enforcing such privacy is impossible under many goals. When relaxing that privacy needs to be protected for all users, delays of elicitation protect privacy of some. Hence, in combination to other goods, privacy becomes a scarce good.

1.2.2 A Personalized Introduction for Economists

The contributions of this dissertation span multiple fields within economics. Chapters 2 and 4 contribute to the literature in theoretical Industrial Organization by analyzing how personalization algorithms influence product demand. Chapters 3 and 6 engage with microeconomic theory and mechanism design, focusing on the strategic implications of personalization and privacy-preserving mechanisms. Finally, Chapter 5 presents an argument for a novel data modality in applied microeconomic research, emphasizing the value of user-generated regret data in evaluating personalized experiences.

Industrial Organization

Chapters 2 and 4 examine how algorithmic selection of *choice architectures* influences market concentration. To formally set up this question, consider a structural model of (random-coefficients, discrete-choice) product demand (S. Berry, Levinsohn, and Pakes 1995; S. Berry, Levinsohn, and Pakes 2004):

$$u_{ijt} = x_{jt\,\top}\beta_i - \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt}. \tag{1.1}$$

Here, i is a consumer making a choice in a market $t \in \mathbb{N}$, and j is an alternative with characteristics x_{jt} and price p_{jt}. The term ξ_{jt} captures unobserved product quality, while ε_{ijt} is an idiosyncratic utility shock. By convention, Greek variables are latent, and Latin variables are observed. We collect the parameters α_i and β_i into a type $\theta_i = (\alpha_i, \beta_i)$ by θ, referring to it as the *consumer's type* throughout this thesis. The behavioral model (1.1) is closed by assuming that in each choice scenario t, consumer i selects:

$$j \in \arg\max_{j \in C_t} u_{ijt},$$

where $C_t \subseteq J$ represents the *consideration set*—our model of a choice architecture—chosen by an algorithm. We will refer to consideration sets also as *personalized experiences*. The selection of consideration sets determines *market shares* s_j for products $j \in J$ and impacts realized *consumer surplus* CS.

Chapters 2 and 4 analyze the market concentration effects of both heuristic and (worst-case) optimal algorithms for selecting consideration sets. We call the distribution of choices of personalized experiences $x \in X$ *algorithmic demand*. These chapters establish conditions under which *noisy observations of consumer types* θ the existence of personalization increases concentration.

To highlight the dependence of mean utility on *consideration sets* (denoted $x \in X$) and *consumer types* ($\theta \in \Theta$), we denote the expected utility function $u(x;\theta)$ where $u(x;\theta)$ represents the mean utility of type θ given consideration set x.

Chapters 2 and 4 explicitly state the assumptions governing the algorithm's *selection of recommendation sets* and the *observations* on which its decisions are based. These insights contribute to understanding how personalization algorithms shape *competitive dynamics and market concentration* through their influence on consumer decisions.

Mechanism Design

The contributions to mechanism design in this thesis, Chapters 3 and 6, focus on the sequential nature of preference elicitation. Both chapters challenge the applicability of the *revelation principle* (compare Myerson (1982)), which asserts that mechanisms can be reduced to direct-revelation forms where agents truthfully report their full preferences. In many personalization settings unrestricted communication is unrealistic due to practical, strategic, or normative requirements. We discuss two settings where practical and normative, respectively, concerns lead to a requirement for sequential elicitation.

Before delving into the contributions, it is helpful to consider how personalization compares to other designed markets. We claim that markets with personalization are characterized by a *high number* of decisions of relatively small importance. Other designed markets involve either individuals for which allocation is highly consequential (school choice Abdulkadiroğlu, Pathak, and Alvin E Roth (2005), medical residency Alvin E. Roth and Peranson (1997), kidney exchange Alvin E. Roth, Sönmez, and Ünver (2004)), or companies (financial markets Budish et al. (2023), markets for pollution permits Joskow, Schmalensee, and Bailey (1998) spectrum reallocation Milgrom and Segal (2020), advertising Edelman, Ostrovsky, and Schwarz (2007)). Participants in such designed markets hence often invest high amounts of cognitive resources, making incentive, computational, and communicational challenges crucial for the successful design of markets.

Personalized experiences, in contrast, involve ordinary people who make frequent decisions, often automatic and with less thought: which video to watch when scrolling on an app, which documentary to watch at night, and which song to play next are settings where less conscious effort is spent, and error is more likely. On the other hand, these choices are so frequent that they become immensely consequential and shape hours of every day for users. The existence of errors, and the high frequency of interactions informs the studies in this thesis on the role of noise, justified elicitation, and novel data modalities.

Chapter 3 examines a setting where personalization is primarily driven by user actions. Instead of directly specifying outcomes after type reports, the mechanism selects a *distribution* over possible outcomes—referred to as *generations*. Users then sample from this distribution at a personal cost. We show that it is possible that the fact that users search can lead to over-representations of minority types.

Chapter 6 formalizes the concept of *contextual integrity*, introduced by Nissenbaum (2004) and further developed by Benthall, Gürses, and Nissenbaum (2017). Contextual integrity posits that all information exchange occurs within a specific context and that privacy violations arise when contextual norms are breached. This principle is also reflected in the European Union's General Data Protection Regulation (GDPR), particularly Article 5(1)(a), which enforces *purpose limitations* on data usage.

The first part of Chapter 6 demonstrates that in many settings, privacy can only be preserved jointly across participants but not on an individual basis, making full contextual privacy infeasible. This necessitates a normative decision on whose privacy to prioritize within the mechanism's design. Privacy hence is scarce, and its allocation is a normative tradeoff for the system designer.

The second part of Chapter 6 introduces strategies for improving privacy in mechanisms. Using classical examples of one-dimensional preference domains, it shows that privacy can be enhanced by limiting information requests to essential cases. This approach leads to mechanisms of the "join" type, where the timing of information elicitation is strategically adjusted—trading off privacy for early versus late participants.

Applied Microeconomics

The final chapter, Chapter 5, introduces an open-source data collection tool designed to capture a novel form of evaluation of personalized experiences: *regret* of past consumption. Using youtube.com as a case study, we demonstrate how regret data can be systematically elicited from user data exports, leveraging rights granted under GDPR's Article 20 on data portability. This approach provides a new empirical modality for studying the long-term effects of algorithmic recommendations on user well-being.

1.3 Previous Publication of Material

This section details prior publications that contributed to this thesis and acknowledges those who provided valuable feedback.

- Chapter 2 is based on A. Haupt and Narayanan (2024). I thank my collaborator Aroon Narayanan, as well as seminar audiences at Harvard, MIT, and Microsoft Research New England. I am also grateful to Phillip J.K. Christoffersen and the MathOverflow user FEDJA for their helpful comments.

- Chapter 3 presents original work that has not previously appeared. It has benefited from discussions with Costis Daskalakis, Wei Zhang, Stewart Slocum, and Dylan Hadfield-Menell.

- The data and analysis in Chapter 4 were originally published in A. Haupt, Podimata, and Hadfield-Menell (2023), co-authored with Chara Podimata and Dylan Hadfield-Menell. I thank my collaborators for their contributions to data collection, as well as Sendhil Mullainathan for his valuable insights.

- The open-source tool described in Chapter 5 is joint work with Mihaela Curmei and was published in A. Haupt and Curmei (2024). This project has benefited from the support of principal investigators Benjamin Recht and Dylan Hadfield-Menell. I also thank audiences at MIT and Harvard University's Berkman Klein Center for Internet and Society, as well as Sendhil Mullainathan, for their feedback.

- Chapter 6 is joint work with Zoë Hitzig and is available online as A. Haupt and Hitzig (2022). It is currently under revise-and-resubmit at the *American Economic Review*. I am deeply grateful to my collaborator Zoë Hitzig. Additionally, we thank Eric Maskin, Shengwu Li, Ben Golub, Yannai Gonczarowski, Jerry Green, Chang Liu, Dirk Bergemann, Stephen Morris, Mohammad Akbarpour, Ilya Segal, Paul Milgrom, Kevin He, Navin Kartik, Ludvig Sinander, Clayton Thomas, Oghuzan Celebi, David Laibson, Modibo Camara, Katrina Ligett, Benji Niswonger, and Max Simchowitz for their valuable discussions. We also acknowledge seminar, workshop, and conference audiences for their comments, including participants at the Harvard and MIT Theory Lunch workshops, seminars at Oxford, Stanford GSB, Boston College, and Boston University, as well as attendees at the Penn State Economic Theory Conference, the Penn/Wharton Mini-Conference in Economic Theory, the Stony Brook Game Theory Festival, and the Berkeley SLMath Workshop on Mechanism and Market Design.

Generative Artificial Intelligence (ChatGPT 4o and OpenAI o1) was used for initial outlines of the introduction and conclusion chapters of this thesis (Section 1.3 and chapter 7) and to create TikZ figures (Figures 4.1 to 4.3).

Chapter 2

Risk Preferences of Learning Algorithms

This first chapter considers a class of algorithms for the choice of personalized experiences and the determinants for the probability with which they choose personalized experiences $x \in X$, which we call *algorithmic demand*. Consider a setting where at each time point, a user with type $\theta_t \sim F \in \Delta(\Theta)$ arrives, a personalization algorithm chooses a personalized experience $x_t \in X$, which the user consumes and derives utility $u(x;\theta)$ from. The personalization algorithm observes a noisy measurement of utility derived from this interaction, $r_t = u(x;\theta) + \varepsilon(x,\theta)$. Here, $\varepsilon(x,\theta)$ is a zero-mean measurement error. The gold standard for this problem is to derive an algorithm selecting x_t based on all observations of $x_1, x_2, \ldots, x_{t-1}$ and $r_1, r_2, \ldots, r_{t-1}$ to find an element that maximizes expected utility $\mathbb{E}_{\theta\sim F}[u(x;\theta)]$ in the limit as $t \to \infty$, which is known as the *no-regret* property in online learning (Lattimore and Szepesvári 2020).

We will show in this chapter that engineering decisions for algorithms within the class of no-regret algorithms have consequences on which personalized experiences are chosen. The well-known ε-Greedy algorithm favors personalized experiences with lower variance in r_t making it what we call a *risk-averse algorithm*. The variance of the feedback r_t can be high due to heterogeneity and measurement error. Indeed, by the law of total variance, we have

$$\mathrm{Var}[r_t] = \mathrm{Var}_{\theta\sim F}[u(x;\theta)] + \mathbb{E}_{\theta\sim F}[\mathrm{Var}[\varepsilon(x;\theta)]].$$

The first term corresponds to heterogeneity that the algorithm cannot condition on. We can think of this as missing information about the intent or ambiguity of a user's query. The latter is the average measurement error. A bias of an algorithm leads to disfavor personalized experiences x that either have high unobserved heterogeneity, high measurement error, or both. The article is descriptive, not prescriptive, but presents two algorithms that do not favor personalized experiences based on the variance of r_t.

2.1 Introduction

Decision-makers often confront the same problem repeatedly, using the outcomes of their choices in the past to guide their actions in the future. For example, credit scores are used to approve or deny credit based on a potential borrower's credit history, and pretrial detention decisions are made based on the defendant's criminal history. Who gets the money to advance their lives—and who gets put in jail with curtailed liberties—will crucially depend on *how* prior data is used to make those decisions.

Many heuristics for solving these types of problems—essentially revolving around keeping an estimate of how well an action has performed in the past—have been developed and formalized into algorithms. They are now widely used in the economy, making decisions from product recommendations to highly consequential areas such as credit approval and pretrial detention.

While most deployed algorithms have a plethora of provable desirable properties, such as identifying the best option in the limit (*no-regret*), we define and demonstrate a bias in widely used learning algorithms: risk aversion, which emerges without being explicitly specified by the algorithm designer. For example, consider a repeated binary choice: Either pull lever A and get a deterministic payoff of 0, or pull lever B and get a stochastic payoff of 1 or -1, with equal probability. At any point in time, contingent on past observations of payoffs from the action taken, an algorithm chooses an action to take next. What is the probability that the learning algorithm selects each action after t rounds of interaction? A risk-averse learning algorithm chooses the deterministic payoff action more often than the other; a risk-neutral learning algorithm chooses them with equal probability. We prove that a classic algorithm, ε-Greedy, is risk-averse. This property holds without ε-Greedy being explicitly designed to be risk-averse. It is an emergent property of the algorithm.

Risk aversion of algorithms is more than a mere intellectual curiosity. It has stark implications across many real-world applications, particularly for fairness in algorithmic choice. This is especially true given that learning algorithms are increasingly being deployed for highly consequential decisions in the form of pretrial algorithms, risk assessment tools (RATs), and credit scoring algorithms, which necessitates extra attention to their emergent and unintended properties. In many of the economic settings where algorithms are deployed to make such decisions, a risk-averse algorithm can perpetuate deep inequities in society. As a concrete example, consider a firm making credit decisions using a risk-averse algorithm. Underrepresented minorities often have wide variances in their credit scores, shaped partly by historic inequities in access to good credit opportunities:

> *As the white suburbs and black inner cities diverged in their mortgage access, two different credit markets emerged in both zones. Lower-risk mortgages led to higher wealth and stability in the white suburbs. These conditions also led to a healthy consumer credit market. In the redlined black ghettos, the economic climate was radically different.* (Baradaran 2019, p. 893)

When faced with minority applicants with higher variability in credit history, a risk-averse algorithm may decide to systematically deny them *even if* it would have approved privileged applicants with similar expected repayment probability but features that are correlated with less variability in credit repayment. They hence perpetuate centuries of iniquity. Yet another setting in which risk aversion can play a significant role is recommendation systems. The choice of recommended products and search results are determined by how valuable the recommendation is deemed to be. Here, too, risk aversion can lead to the recommendation system suppressing "noisier" content—which, in most cases, will be the less mainstream, more marginalized content—even when its deployers do not find such bias desirable. In the long run, this bias can also lead to the homogenization of the content on these platforms, as more divergent content is not recommended as frequently by the algorithm.

Our first formal result is that the ε-Greedy algorithm exhibits risk aversion, preferring deterministic over non-deterministic actions of the same variance. The intuition for the emergence of risk aversion lies in the way ε-Greedy estimates the payoff of each action. If an algorithm's estimate of each action is the simple average of the observed payoffs in the past from this action, which is what ε-Greedy does, its estimates will be biased because the algorithm undersamples actions after low payoff realizations. We then discuss two corrections to the algorithm that enable it to be risk-neutral. Our first correction, which we call the *Reweighted ε-Greedy*, counters the undersampling propensity by adjusting its estimate to account for the probability with which an action was chosen. We show that Reweighted ε-Greedy is risk neutral. We also propose another correction for a broader class of settings: the *Optimistic ε-Greedy*. It adds an optimism term to the estimate that corrects bias asymptotically, see (Auer 2003; Auer, Cesa-Bianchi, and Fischer 2002; Lattimore and Szepesvári 2020). Our third formal result shows that this correction also makes ε-Greedy risk neutral. We use simulations to explore the necessity of conditions we make in our theoretical analysis and the transient persistence of risk behavior even with unequal expected values for the actions.

2.1.1 Related literature

There is a large literature on learning from feedback in economics. We highlight the papers Bolton and Harris (1999), Keller, Rady, and Cripps (2005), Klein and Rady (2011), Baek and Farias (2021) as theoretical contributions, Bergemann and Välimäki (2018) is a survey. The paper Bardhi, Y. Guo, and Strulovici (2020) demonstrates that even arbitrarily small differences in early-career discrimination can be highly consequential later in life. Our results complement this literature by showing that algorithmic learning can exhibit unintended discrimination with strong consequences in the long run. A second branch of literature studies learning by economic agents in empirical settings. Farber and Gibbons (1996), Altonji and Pierret (2001), and Baek and Makhdoumi (2023) study learning by employers, providing testable predictions for wage dynamics. Crawford and Shum (2005) applies learning to demand for pharmaceutical drugs. Recently several papers have considered the behavior of learning algorithms in simulations, particularly in relation to collusion. Calvano et al. (2020), Musolff (2022), Brown and MacKay (2023), and Banchio and Mantegazza (2023) find in different game-theoretic settings that pricing algorithms learn to play collusive equilibria, raising antitrust concerns about the use of such algorithms in pricing. Banchio and Mantegazza (2023) shows that in games, spontaneous collusion can arise because of correlation of play and asymmetric sampling. The intuition presented here relies on a similar causal channel—asymmetric sampling—but considers different algorithm classes and stresses synchronization as opposed to distributional questions. It also considers a setting with a single algorithm making decisions, as opposed to multiple algorithms interacting in a game.

Our paper is connected to the computer science literature on the effect of biased payoff estimates in recommendation systems. Marlin and Zemel (2009) observes that online learning in recommendation systems leads to confounding of average user scores in recommendation systems and proposes algorithmic interventions to correct this bias. Chaney, Stewart, and Engelhardt (2018) proposes a model of recommendation and shows that recommendation systems' biased estimates of user preferences can increase homogeneity and decrease user utility. Our study focuses on the effect of noise and the propensity of taking particular actions and does not directly consider the bias in estimates.

We also relate to the study of fairness in bandit problems. While Joseph et al. (2016) considers fairness (which is a finite-time variant of our notion of risk neutrality) as a constraint for algorithm design and constructs algorithms that approximately satisfy it, this paper provides evidence on the risk preferences of an existing algorithm, ε-Greedy, and proposes two ways to mitigate risk preferences; see also Patil et al. (2021) and Y. Liu et al. (2017) for treatments of fairness in bandit problems. Dai et al. (2024) considers the relationship of sampling rates to the regret of algorithms and provides improved regret bounds for algorithms that are sampling actions in a more balanced fashion.

Our work also relates to the study of notions of "rationality" for algorithms (Raman et al. 2024; Rahwan et al. 2019). This literature aims to understand in which environments algorithms behave according to behavioral axioms that were developed for humans.

Finally, we relate to the regret analysis of bandit algorithms under diffusion scaling. Kalvit and Zeevi (2021a) studies this for the Upper Confidence Bound algorithm (compare Theorem 2.3). Fan and Glynn (2024) derives the limit action distribution of Thompson sampling as a solution to a random ordinary differential equation.

2.1.2 Outline

The structure of the rest of this paper is as follows. In Section 2.2, we introduce our online learning setup and our definition of risk aversion, along with formal definitions of our algorithms. The result on ε-Greedy's risk aversion is presented in Section 2.3. We discuss two corrections of risk aversion in Section 2.4. We complement our theoretical analysis with simulations in Section 2.5. We conclude in Section 2.6. An appendix contains additional simulations.

2.2 Model

In a bandit problem, a decision maker repeatedly takes an action from a finite set A, $|A| < \infty$. Each action a is associated to a sub-Gaussian distribution $F_a \in \Delta(\mathbb{R}), a \in A$ with expectation μ_a and variance proxy σ_a^2.[1] An algorithm π generates (potentially random) sequences of *actions* $(a_t)_{t\in\mathbb{N}}$ and *rewards* or *payoffs* $(r_t)_{t\in\mathbb{N}}$. For each $t \in \mathbb{N}$, repeatedly, the algorithm chooses an action $a_t \sim \pi_t$ and receives a reward $r_t \sim F_{a_t}$. That is, an *algorithm* is a function $\pi\colon \bigcup_{t=1}^{\infty}(A \times \mathbb{R})^t \to \Delta(A)$. We denote action-reward histories by $(a_{1:t}, r_{1:t})$ and the probability that action $a \in A$ is chosen in round t by $\pi_{at} = \pi(a_{1:t}, r_{1:t})_a$. Denote $N_a(t) := |\{1 \leq t' \leq t : a_t = a\}|$ the number of times action a has been chosen up to time t.

[1] A distribution is sub-Gaussian if $\int e^{\lambda X}\,\mathrm{d}F_a(x) \leq \exp(\frac{\lambda^2\sigma_a^2}{2})$. In this case, σ_a^2 is called the *variance proxy*.

The main concept in this paper is a notion of *risk aversion* of algorithms. An algorithm is risk-neutral if it chooses (asymptotically) uniformly from amongst actions of equal expectation. In the long run, risk-averse algorithms prefer less risky actions than others of the same expectation. In the extreme case where the algorithm exclusively chooses (asymptotically) the least risky action among those of the same expectation, we call them *perfectly* risk averse.

Definition 2.1. We call π *risk-neutral* if for any actions $a, a' \in A$ such that $\mu_a = \mu_{a'}$,

$$\lim_{t \to \infty} \mathbb{P}[a_t = a] = \lim_{t \to \infty} \mathbb{P}[a_t = a'].$$

We call an algorithm *risk-averse* if for all $a, a' \in A$ such that $\mu_a = \mu_{a'}$ and $F_{a'} \prec_{\mathrm{SOSD}} F_a$, it holds that[2]

$$\lim_{t \to \infty} \mathbb{P}[a_t = a] > \lim_{t \to \infty} \mathbb{P}[a_t = a'].$$

An algorithm is *perfectly risk averse* if for any instance for which there is $a \in A$ such that either $\mu_{a'} < \mu_a$ or $F_{a'} \prec_{\mathrm{SOSD}} F_a$ for all $a' \in A \setminus \{a\}$,

$$\lim_{t \to \infty} \mathbb{P}[a_t = a] = 1.$$

This paper considers the ε-Greedy algorithm and two variants of ε-Greedy with different statistics.

Definition 2.2 (ε-Greedy)**.** Let $(\varepsilon_t)_{t \in \mathbb{N}}$ be a $[0, 1]$-valued sequence. ε-Greedy chooses the empirically best action with probability $1 - \varepsilon_t$, and randomizes between all the actions with probability ε_t, *i.e.*

$$\pi_{at} = \begin{cases} \mathrm{Unif}(\arg\max_{a \in A} \mu_a(t-1)) & \text{w.p. } 1 - \varepsilon_t \\ \mathrm{Unif}(A) & \text{w.p. } \varepsilon_t, \end{cases}$$

where $\mu_a(t-1)$ is historical average payoff

$$\mu_a(t) := \frac{1}{N_a(t)} \sum_{\substack{1 \le t' \le t \\ a_{t'} = a}} r_{t'}.$$

When ε-Greedy takes an action to maximize $\mu_a(t-1)$, we say it *exploits* or *takes an exploitation step*. Otherwise, it *explores* or *takes an exploration step*. We also call $\mu_a(t-1)$ ε-Greedy's *statistic*.

The first variant reweighs data points to change their importance.

[2] Distribution F dominates F' in second-order stochastic dominance, $F \geq_{\mathrm{SOSD}} F'$ if $\int u \, \mathrm{d}F \geq \int u \, \mathrm{d}F'$ for all concave, non-decreasing functions u, with a strict inequality for some such function u.

Definition 2.3 (Reweighted ε-Greedy)**.** Reweighted ε-Greedy uses a reweighted payoff estimate as a statistic:

$$\mu_{a,r}(t) = \frac{1}{N_a(t)} \sum_{\substack{1 \le t' \le t \\ a_t = a}} \frac{r_{t'}}{\sqrt{\pi_{at'}}}.$$

A second intervention adds an optimism term to the statistic of ε-Greedy.

Definition 2.4 (Optimistic ε-Greedy)**.** Optimistic ε-Greedy adds an optimism term to its statistic:

$$\mu_{a,o}(t) = \mu_a(t) + \rho\sqrt{\frac{\log(t)}{N_a(t)}}.$$

If action a has not been chosen until time t, $\mu_{a,o}(t) = \infty$.

2.3 Risk aversion of ε-Greedy

We first show that ε-Greedy is perfectly risk-averse.

Theorem 2.1. *Let $(\varepsilon_t)_{t\in\mathbb{N}}$ such that $\varepsilon_t \to 0$ and $\sum_{t=1}^{\infty} \varepsilon_t = \infty$. If there is a deterministic, centered dominant action, and all other actions have symmetric continuous distributions, ε-Greedy is perfectly risk-averse.*

A discussion on the conditions in this result is in order before we move on to the proof. Both hypotheses on the exploration rates are necessary to yield a *no-regret algorithm*, compare Lattimore and Szepesvári (2020), and are standard in the literature. We show in our simulations in Section 2.5 that relaxing the requirement of determinism of a dominant action leads to risk aversion, but not perfect risk aversion, as does relaxing the symmetry requirement.

The intuition behind this result lies in the sampling bias of ε-Greedy. Upon receiving a low payoff realization for an action, it becomes less likely to choose that action and, hence, less likely to receive data to correct its estimate. This means that it keeps a pessimistic estimate of reward. In contrast, for a high payoff realization, the algorithm frequently samples this action, leading the algorithm to correct its estimate.

Proof of Theorem 2.1. We first observe that we can restrict to bandit problems of two actions with reward distributions of the same expectation, one deterministic dominant action a and another action a' with a continuous, symmetric distribution. Under the assumptions on the exploration rate made in the theorem, ε-Greedy chooses dominated actions with vanishing probability. We prove this in Lemma 2.1 in the appendix. As this probability is low, we may consider instances of actions of equal expectation. In addition, a union bound shows that if the probability that the algorithm chooses action a over any single action a' converges to 1, this implies that this action will be chosen with probability one among all actions of the same probability. Hence, it is without loss to restrict to two-action bandit problems.

We also observe that the result is trivial for two deterministic actions with the same expectation. Hence, we may assume that $A = \{a, a'\}$, $\mathrm{Var}(F_a)$ has a positive variance, and $F_{a'} = \mathbb{1}_{\{0\}}$. Furthermore, it is without loss to assume that the deterministic action is centered: ε-Greedy is invariant to the addition of a constant to all reward distributions As a final reduction step, as $\varepsilon_t \to 0$, it is sufficient to show that it becomes unlikely that ε-Greedy chooses a' in exploitation steps, or

$$\mathbb{P}[\mu_a(t) < \mu_{a'}(t)] \to 0.$$

We express this event as a property of a stochastic process. The sum of the payoffs of the non-deterministic action, denoted by $(X_t)_{t\in\mathbb{N}}$, is a sufficient statistic for the dynamics of the algorithm. $(X_t)_{t\in\mathbb{N}}$ is a lazy random walk starting at the origin, $X_0 = 0$, with transition kernel

$$X_{t+1} = \begin{cases} X_t + r & \text{with probability } (1 - \frac{\varepsilon_t}{2})\mathbb{1}_{X_t>0} + \frac{1}{2}\mathbb{1}_{X_t=0} + \frac{\varepsilon_t}{2}\mathbb{1}_{X_t<0}, \\ X_t & \text{else,} \end{cases}$$

where $r \sim F_a$. We call this the *advantage walk* and depict it in Figure 2.1. Consider the time since the last time that the random walk crossed zero, $\tau_t := t - \max\{1 \leq t' \leq t | X_{t'-1} \leq 0 \leq X_{t'}\}$. We claim that $\tau_t \to \infty$ as $t \to \infty$, in probability.

To prove this claim, it suffices to show that for any $c \geq 0$ and $\varepsilon > 0$, there is t' such that $\mathbb{P}[\tau_{t'} \leq c] \leq \varepsilon$ for all $t' \geq t$. Observe that for any $c \geq 0$, there is $C > 0$ and $t \in \mathbb{N}$ such that for all $t' \geq t$

$$\mathbb{P}[\tau_{t'} \leq c] \leq \mathbb{P}[|X_{t'-c}| < C] + \frac{\varepsilon}{2} \leq \varepsilon.$$

The first inequality is a consequence of the sub-Gaussianity of F_a. The second inequality is a result of $\mathrm{Var}(F_a) > 0$, the conditional independence of increments, and $\sum_{t'=1}^{t-c} \varepsilon_{t'} \to \infty$ as $t \to \infty$.

We also define the distribution of the number of steps taken since τ_t on the positive side. These are distributed as

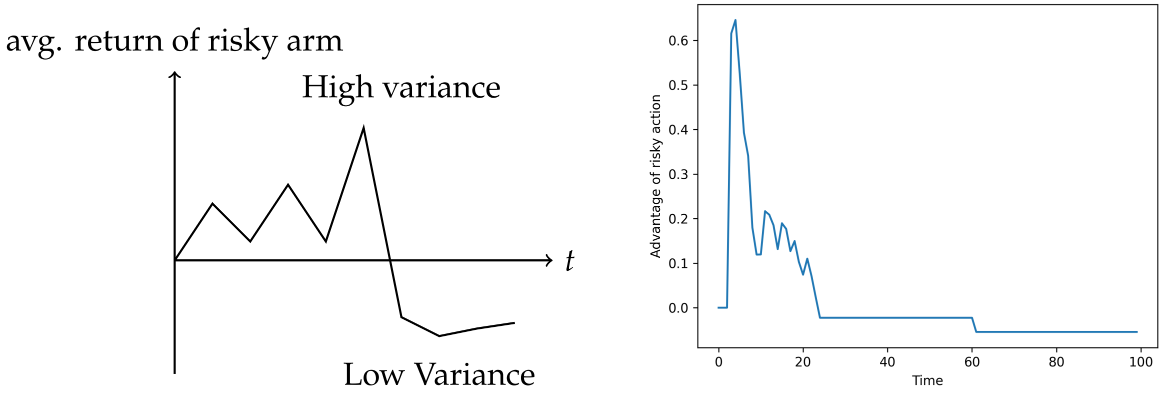
$$P_t \sim \sum_{t'=t-\tau_t}^{t} Z_{t'}, \quad Z_t \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(1 - \varepsilon_t/2).$$

As P_t is a sum of i.i.d. random variables, by Hoeffding's inequality, P_t is close to its expectation $\mathbb{E}[P_t|\tau_t]$ for large enough τ_t. Conditioning on τ_t,

$$\mathbb{E}[P_{\tau_t}|\tau_t] = \sum_{t'=t-\tau_t}^{t} 1 - \frac{\varepsilon_{t'}}{2}.$$

We have

$$\mathbb{P}[X_t > 0] \leq c\mathbb{P}[Y_1, Y_2, \dots, Y_{P_t} > 0],$$



(a) Illustration of the advantage walk

(b) One realization of the advantage walk for ε-Greedy where the safe action has distribution $\mathbb{1}_{\{0\}}$ while the risky action has distribution $U[-1,1]$

Figure 2.1: The advantage walk for ε-Greedy. The main intuition for risk aversion in online algorithms is that a random walk with non-uniform variance spends more time in places with lower variance.

where $Y_0 = 0$ and $(Y_t)_{t\in\mathbb{N}}$ is a standard random walk with increment distribution F_a. For this inequality, we can choose $c \geq 1/\mathbb{P}[r > X_{t-\tau_t}]$, where $r \sim F_a$ is independent of $(X_t)_{t\in\mathbb{N}}$. This follows as a single step from zero could lead from 0 to $X_{t-\tau_t}$, or a higher value. Because $X_{t'-\tau_{t'}}$ is reached from $X_{t'-\tau_{t'}-1} < 0$, $\mathbb{P}[r > X_{t'-\tau_{t'}}] > 0$ must be positive, and hence c is well-defined.

Hence, for any $\delta > 0$, there is $t' \in \mathbb{N}$ such that for all $t \geq t'$,

$$\mathbb{P}[X_t > 0|\tau_t] \leq \frac{c}{\sqrt{\pi(\sum_{t=t-\tau_t}^{t} 1 - \frac{\varepsilon_{t'}}{2})}}(1+\delta). \tag{2.1}$$

This inequality uses the well-known property that the probability of a random walk stays positive until time t with probability approximately $\frac{1}{\sqrt{\pi t}}$ (U. Frisch and H. Frisch 1995, Eqn. 35). (This property also holds for Rademacher-distributed increments and is the only place where we use the assumption that our distribution is continuous and symmetric.)

Observe that (2.1) approaches 0 as $\tau_t \to \infty$ and recall that $\tau_t \to \infty$ as $t \to \infty$, in probability. Given these two facts,

$$\mathbb{P}[X_t > 0] \xrightarrow[t\to\infty]{\mathbb{P}} 0,$$

which concludes the proof. □

We highlight that a similar argument applies to actions that are dominated by others in expectation. For two actions a, a' of the same expectation, such that a is deterministic but a' is not, and a third action a'' such that $\mu_{a''} > \mu_a, \mu_{a'}$, the probability that a is taken conditional a or a' are taken converges to one.

2.4 Achieving risk neutrality

Next, we propose variants to the ε-Greedy statistic to achieve risk neutrality. These corrections are motivated by the technical analysis of the previous theorem as well as known algorithmic ideas (*e.g.*, Auer (2003)). Our corrections highlight the mechanisms leading to bias, and their redressal.

2.4.1 A reweighting approach to risk neutrality

The first approach stems from our analysis of variance: we should reweight data to achieve risk neutrality. A reweighted ε-Greedy provably is risk-neutral if exploration is sufficiently high.

Theorem 2.2 (Reweighted ε-Greedy)**.** *Let $\varepsilon_t = t^{\frac{1}{2}+\kappa}, \kappa \in (0, \frac{1}{2})$. Reweighted ε-Greedy is risk-neutral for two centered actions, one of which is deterministic.*

The reweighting proposed here means that payoffs resulting from currently unfavored actions are weighted more highly in the statistic of the algorithm. In the credit scoring setting, if the option of rejecting the loan is currently favored, the outcome of any credit that is given (due to exploration) is weighted more highly in the statistic. Thus, this correction tells us that we should assign more importance to outcomes that resulted from choices that seemed apriori less attractive.

It is worth noting that this reweighting does *not* lead to an unbiased estimator of action rewards. We show in simulations in an appendix that an unbiased estimator leads to a risk-loving algorithm, Section 2.B. This is a result of what the rescaling does to the algorithm's statistic: The goal of the algorithm proposed here is to equalize variance, which is at odds with producing an unbiased estimator.

Several comments on the theorem are in order. The first assumption on exploration means that a sufficient amount of exploration is needed for this algorithm to be risk-neutral. We provide a simulation in Section 2.5 showing that this condition is important for risk neutrality. The assumption of centeredness and determinism is needed for our proof, but risk aversion seems to hold beyond them, as we show in Section 2.5. On the other hand, the requirement that there are only two actions with the same expectation is crucial, as we show in another simulation in Section 2.5. In an appendix, we discuss the regret properties of this algorithm.

Proof. Note that if both actions are deterministic, the conclusion of the theorem holds trivially. Otherwise, denote the non-deterministic action by $a \in A$. The evolution of choices can be expressed only based on the stochastic process

$$Y_t = \sum_{\substack{1 \leq t' \leq t \\ a_t = a}} \frac{r_t}{\sqrt{\pi_t}}.$$

If $Y_t > 0$, action a is chosen with probability $1 - \varepsilon_t/2$, if $Y_t = 0$, it is chosen with probability $1/2$, and if $Y_t < 0$, then it is chosen with probability $\varepsilon/2$. We define the random array

$$X_{t't} = \frac{1}{\sqrt{t}} Y_{t'}$$

This random array has the following properties:

Martingale It is an L^2-martingale array, *i.e.* $(X_{t't})_{1 \leq t' \leq t}$ is a square-integrable martingale with respect to its natural filtration.

Asymptotic Variance The conditional variances of martingale increments are constant (and deterministic).

$$\begin{aligned}
\sum_{t'=1}^{t} \mathbb{E}[(X_{t't} - X_{(t'-1)t})^2 | X_{(t'-1)t}] &= \sum_{t'=1}^{t} \sum_{a \in A} \pi_{a(t')} \mathbb{E}\left[\frac{r_a^2}{\sqrt{t\pi_{a(t')}}^2} \middle| X_{(t'-1)t} \right] \\
&= \sum_{t'=1}^{t} \frac{1}{t} \sum_{a \in A} \sigma_a^2 \\
&= \sigma_a^2.
\end{aligned}$$

In particular, as $n \to \infty$, $\mathbb{E}[(X_{t't} - X_{(t'-1)t})^2 | X_{(t'-1)t}] \to \sigma_a^2$ in probability.

Lindeberg Condition For any $\varepsilon > 0$, we have that

$$\begin{aligned}
\sum_{t'=1}^{t} \mathbb{E}[(X_{t't} - X_{(t'-1)t})^2 \mathbb{1}_{|X_{t't} - X_{(t'-1)t}| \geq \varepsilon} | X_{(t'-1)t}] &\leq 2 \sum_{t'=1}^{t} \frac{(t')^{1-2\kappa}}{t} \mathbb{E}[r^2 \mathbb{1}_{|r| \geq t^{\frac{1}{2}} (t')^{\frac{1}{2}-\kappa} \varepsilon}] \\
&\leq \frac{2}{t} \sum_{t'=1}^{t} (t')^{1-2\kappa} \sigma_a^2 e^{\frac{\lambda^2 \sigma_a^2}{2} - \lambda t^{\frac{1}{2}} (t')^{\frac{1}{2}-\kappa} \varepsilon} \\
&\leq \frac{2}{t} \sum_{t'=1}^{t} t^{1-2\kappa} \sigma_a^2 e^{\frac{\lambda^2 \sigma_a^2}{2} - \lambda t^{\frac{1}{2}} \varepsilon} \\
&\to 0.
\end{aligned}$$

The first inequality plugs in definitions. The second uses a Chernoff bound. The last uses $1 \leq t \leq t'$. The convergence follows as exponential decay dominates polynomial growth and as convergence of a sequence implies convergence of the Cesàro mean.

Given these conditions, we can apply a Martingale Central Limit Theorem (Hall and Heyde 1980, Corollary 3.1) and conclude that the distribution of X_{tt} converges to $N(0, \sigma_a^2)$. This means that

$$\mathbb{P}[X_{tt} > 0], \mathbb{P}[X_{tt} < 0] \xrightarrow{t\to\infty} \frac{1}{2},$$

and hence

$$\mathbb{P}[a_t = a] = \frac{1}{2}. \qquad \square$$

While this algorithm works for two actions, another approach to risk neutrality, *optimism*, allows us to guarantee risk neutrality for an arbitrary number of actions of equal expectation. We discuss in the next subsection.

2.4.2 An optimism approach to risk neutrality

Another way to modify the statistic is not to reweight but to explicitly favor alternatives that have not previously been chosen as frequently in the past. Conventionally, this is referred to as *optimism* in the multi-armed bandit literature, compare Slivkins (2019, Section 1.3.3). We show that it ensures risk neutrality.

Theorem 2.3 (Optimistic ε-Greedy)**.** *There exists $\rho_0 > 1$ such that for any $\rho \geq \rho_0$ and any $(\varepsilon_t)_{t\in\mathbb{N}}$ with $\varepsilon_t \to 0$, Optimistic ε-Greedy is risk-neutral.*

Proof Sketch. Note first that for exploitation steps of Optimistic ε-Greedy, the policy is the same as Upper Confidence Bound with exploration coefficient ρ, compare Auer (2003). We adapt the proof of Kalvit and Zeevi (2021b, Theorem 2) for our variant of ε-Greedy. Theorem 2 in Kalvit and Zeevi (2021b) shows that an optimistic policy without exploration has the property that

$$\lim_{t\to\infty} \mathbb{P}[a_t = a] \to \frac{1}{|\arg\max_{a\in A} \mu_a|}$$

for all a such that $a \in \arg\max_{a\in A} \mu_a$ and $\rho > \rho_0$. As Optimistic ε-Greedy does not incur regret as we show in Section 2.A, this property implies that the algorithm does not incur regret.

The full proof can be found in Kalvit and Zeevi (2021a, Appendix D). The proof goes as follows. First, show that $N_i(t)/t > 1/2|\arg\max_{a\in A} \mu_a|$ with probability approaching 1 in the limit, *i.e.* the fraction of times any action is chosen can be bounded below in probability. This is proved by means of a union bound and a Hoeffding bound. The operative equation that gets to this lower bound is Kalvit and Zeevi (2021a, Eqn. 40), which depends on Kalvit and Zeevi (2021a, Eqns. 35 and 39). Using this lower bound, we show that $|N_i(t) - N_j(t)|$ is small, *i.e.* the difference in the number of times any two actions are chosen is small as the time goes to infinity. This again uses Markov's inequality, along with the Law of the Iterated Logarithm. Now, consider optimistic ε-Greedy. Since it implements the Upper Confidence Bound policy with probability $1 - \varepsilon_t$ and randomizes otherwise, in either case, it must be eventually choosing uniformly from amongst the highest mean actions, except for the vanishing probability with which it chooses dominated actions. □

It is worth noting why the mathematical intuition from the earlier result on ε-Greedy breaks. In this proof, the main object was a random walk with different variances for positive and negative values of the statistic. Optimism may be seen as introducing a drift towards the origin. The proof shows that this drift is strong enough to correct risk aversion.

This result demonstrates that one way to build a fairer world is with a particular type of optimism. Linking back to the credit decisions example we referenced in the introduction, credit scoring algorithms should evaluate applicants in the best possible light, adjusting for the risk profile of minority applicants by accounting for the fact that there might be less information on them.

2.5 Experiments

For the final section of this paper, we use experiments to confirm our theoretical results and investigate how far our results extend beyond their conditions. Unless noted otherwise, our experiments consider $\varepsilon_t = t^{-1}$, and report confidence bands that are Gaussian 90 % confidence intervals from 100 independent runs.

2.5.1 On risk aversion of ε-Greedy

Our initial set of experiments relate to the conditions in Theorem 2.1. The first experiment considers a setting that satisfies the conditions of Theorem 2.1. We simulate an ε-Greedy algorithm for an instance with three actions—a safe action that has distribution $\mathbb{1}_{\{0\}}$, a riskier action that has payoff distribution $U[-0.5, 0.5]$ while the riskiest action has payoff distribution $U[-1, 1]$. The results of this experiment can be found in Figure 2.2a ε-Greedy converges very quickly to selecting only the safe action.

The next experiment investigates a setting where arm rewards are not symmetric. ε-Greedy chooses between a safe action that has distribution $\mathbb{1}_{\{10\}}$ and a risky action that has distribution an exponential distribution with rate 10 *i.e.* Exp(10). The results are in Figure 2.2b. Even for the asymmetric exponential distribution, we observe perfectly risk-averse behavior.

Next, we consider settings where reward distributions are not centered around 0. We do this by setting the safe action to have distribution $\mathbb{1}_{\{0.5\}}$ while the risky action has distribution $U[2, -1]$. The results from this experiment are in Figure 2.2c. Again, we find that the safe arm is chosen with high probability.

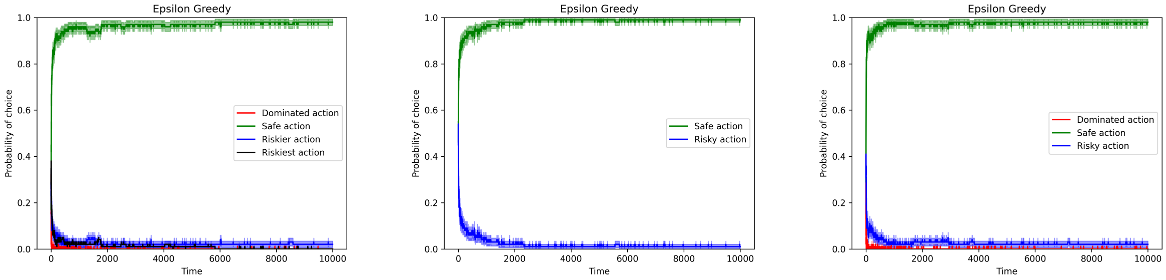
In all of the experiments so far, we had a perfectly safe action with constant reward. Our next experiment explores what happens when there is no optimal deterministic action. We consider an instance with a dominated action with reward distribution $\mathbb{1}_{\{-1\}}$, a safer action with distribution $U[-0.25, 0.25]$, a riskier action to have payoff distribution $U[-0.5, 0.5]$, and the riskiest action to have payoff distribution $U[-1, 1]$. The results are in Figure 2.3a. We find that ε-Greedy still chooses the safer action with significantly higher probability than the riskier one, which in turn is chosen with significantly higher probability than the riskiest one.

An important natural question that arises here is whether emergent risk aversion has implication beyond actions with equal mean payoffs. We show in Figure 2.3b that ε-Greedy's risk aversion has large transient effects before asymptotic guarantees kick in. This clarifies that the bias we identify can persist for a long time—for example, credit decisions can continue to be significantly discriminatory with such an algorithm *even if* the minority candidate has a strictly higher likelihood of repaying the loan.

2.5.2 On risk neutrality of Reweighted ε-Greedy

Our second set of experiments consider Reweighted ε-Greedy. To explore the limits of Theorem 2.2, we run simulations that vary exploration rate and the number of actions. We find that Reweighted ε-Greedy is risk-neutral for two actions across different reward distributions as long as the exploration rate is sufficiently high.

Our first experiment chooses a higher exploration rate than before ($\varepsilon_t = t^{-0.49}$) as required by Theorem 2.2. The instance has a safe action with payoff distribution $\mathbb{1}_{\{0\}}$ and a risky action with payoff distribution $U[-1, 1]$.



(a) Perfect risk aversion.

(b) Perfect risk aversion.

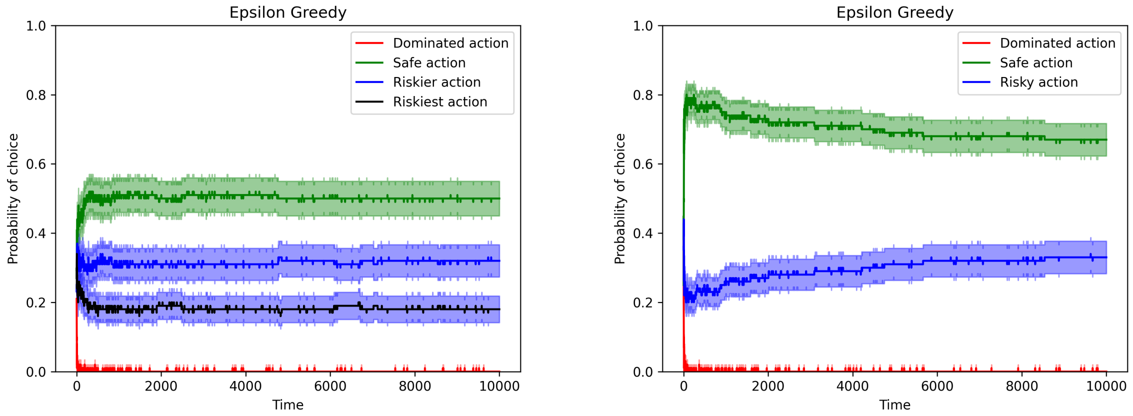
(c) Perfect risk aversion.

Figure 2.2: Plots of the behaviour of ε-Greedy, under and outside the conditions of Theorem 2.1. (a) Three actions with the same mean, ordered in second-order stochastic dominance, and including a deterministic arm. The deterministic arm is chosen most of the time. (b) Two arms with the same expectation, one determinstic, and one having an asymmetric distribution (exponential). While outside of the conditions of Theorem 2.1, the safe action is chosen most of the time. (c) Non-centered arms, ordered in second-order stochastic dominance. While outside of the conditions of Theorem 2.1, the safe action is chosen most of the time.

The second experiment considers again more exploration $\varepsilon_t = t^{-0.49}$. We consider runs of the algorithm on an instance with a safer action with reward distribution $U[0.25, 0.75]$ and a risker action with payoff distribution $U[0, 1]$.

The third instance considers the same instance as the first, but with an exploration rate of $\varepsilon_t = t^{-1}$. Figure 2.4a, Figure 2.4b, and Figure 2.4c respectively show the results. They show that the conclusions of Theorem 2.2 extend beyond the conditions of the Theorem as long as exploration is sufficiently high.

However, the restriction to two actions is crucial. To show this, we run a simulation of Reweighted ε-Greedy where we add a third action with distribution $\mathbb{1}_{\{-1\}}$ to the setup



(a) Risk aversion

(b) Transient risk aversion

Figure 2.3: Results of additional experiments investigating the generality of ε-Greedy's risk aversion. (a) For three centered arms that are ordered in second-order stochastic dominance, the safe arm is chosen most of the time. (b) For a dominated safe action, after 1,000 steps, the safe action is still chosen with probability more than a half.

of where the safe action has payoff distribution $\mathbb{1}_{\{0\}}$ while the risky action has payoff distribution $U[-1,1]$ and set $\varepsilon_t = t^{-0.49}$. Figure 2.5 shows that risk neutrality need not hold in such a setting.

2.5.3 On risk neutrality of Optimistic ε-Greedy

Finally, we consider the assumptions of Theorem 2.3. We do this by running three simulations. In the first one, we consider a high exploration coefficient $\rho = 2$. We consider the behavior of the algorithm on an instance with payoff distributions $U[-0.25, 0.25]$ and $U[-1,1]$, and a dominated action with reward $\mathbb{1}_{\{-1\}}$.

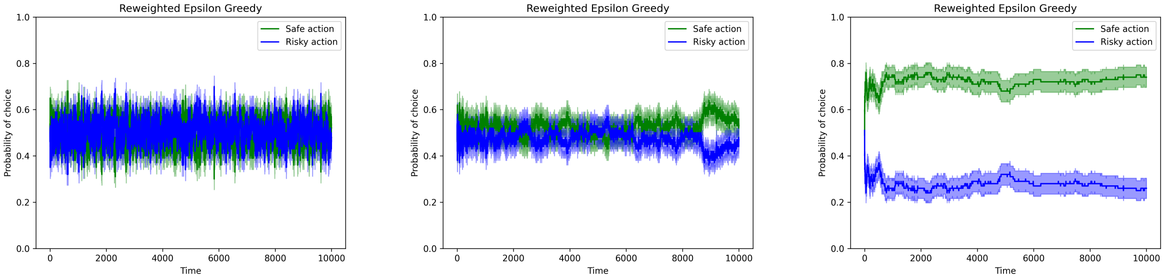
The second experiment considers a high exploration coefficient $\rho = 2$ once more, but an instance consisting of a non-centered reward distributions $U[0.25, 0.75]$, $U[0,1]$, and a dominated action with reward $\mathbb{1}_{\{-1\}}$.

The final experiments consider a lower exploration coefficient $\rho = 0.02$ and an instance with reward distributions $\mathbb{1}_{\{0\}}$ and $U[-1,1]$, and a dominated action $\mathbb{1}_{\{-1\}}$.

The results are in Figure 2.6a, Figure 2.6b, and Figure 2.6c, respectively. While the first two experiments show risk neutrality, the last experiment shows the necessity of a sufficient optimism coefficient for the conclusions in Theorem 2.3 to hold.

2.6 Conclusion

Learning algorithms can have unintended emergent risk behavior, leading to outcomes that may be at odds with the objectives of those who deploy them. The basic intuition behind this is that exploration policies often use simple statistics such as the mean to keep track of the estimated value of each option while not using the fact that other properties



(a) Risk neutrality (b) Risk neutrality (c) Risk aversion

Figure 2.4: Plots showing that the Reweighted ε-Greedy is risk-neutral for two actions as long as exploration is sufficiently high. (a) For the case of two arms that are ordered in second-order stochastic dominance, one of which is deterministic, both arms are chosen equally often, as predicted by Theorem 2.2. (b) While outside of the conditions of Theorem 2.2, this seems to continue to hold for distributions that are ordered in second-order stochastic dominance but no action that is deterministic. (c) for less exploration, Reweighted ε-Greedy fails to choose the arms with the same probability, favouring the safer action.

of their data, such as noise, can affect their learning process. As a consequence, higher variance actions can end up being shunned purely because they yield a bad outcome early on. Corrections to the algorithm's statistic can reinstate risk neutrality.

2.A Regret properties of algorithms

This section provides evidence that the algorithms we consider do not incur regret.

Definition 2.5. An algorithm π *is no-regret* or *does not incur regret* if for all instances such that F_a is sub-Gaussian for all $a \in A$ and for all $a, a' \in A$, $\mu_a < \mu_{a'}$, we have

$$\mathbb{P}[a_t = a] \xrightarrow{t\to\infty} 0.$$

We provide a proof sketch for the following statement, which implies that ε-Greedy does not incur regret.

Lemma 2.1. *Let $\varepsilon_t \to 0$, $\sum_{t=1}^{\infty} \varepsilon_t = \infty$. Also let $a \in A$, $\mu_a < \max_{a \in A} \mu_a$. Then, for any $\delta > 0$, there is $t \in \mathbb{N}$ such that for all $t' \geq t$,*

$$\pi_{at'} \leq \delta.$$

Proof Sketch. Choose $t' \in \mathbb{N}$ such that for all $\tilde{t} \geq t'$, $\varepsilon_t \leq \delta/2$. That is, the probability of exploration steps is small. It is sufficient to show that for some t'' and $\tilde{t} \geq t''$, in exploitation steps,

$$\mathbb{P}[\mu_a(\tilde{t}) - \mu_{a'}(\tilde{t}) \geq 0] \leq \frac{\delta}{2}.$$

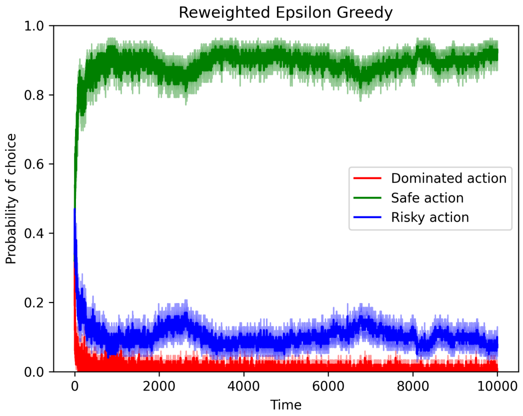


Figure 2.5: Reweighted ε-Greedy with a third dominated arm. The addition of a third arm leads the safe arm to be chosen more frequently.

(a) Risk neutrality (b) Risk neutrality (c) Risk aversion

Figure 2.6: Plots showing that the Optimistic ε-Greedy is quite generally risk-neutral, as long as exploration coefficient ρ is sufficiently high. (a) For two centered arms and a dominated arm, one of which is deterministic, we find that the safer action is chosen most of the time. (b) Also, for two non-centered arms that are ordered in second-order stochastic dominance and a dominated arm, the safe arm is chosen most of the time. (c) For a lower exploration rate and two centered arms ordered in second-order stochastic dominance, one of which is deterministic, risk neutrality fails to hold.

By Hoeffding's inequality, with high probability in $\tilde{t}$, both actions have been chosen at least $\frac{1}{3}\sum_{t=1}^{t''}\varepsilon_t$ times. Conditional on this event,

$$\mathbb{P}[\mu_a(\tilde{t}) - \mu_{a'}(\tilde{t}) \geq 0] \xrightarrow[t\to\infty]{} 0.$$

In particular,

$$\mathbb{P}[\mu_a(\tilde{t}) - \mu_{a'}(\tilde{t}) \geq 0] \leq \frac{\delta}{2},$$

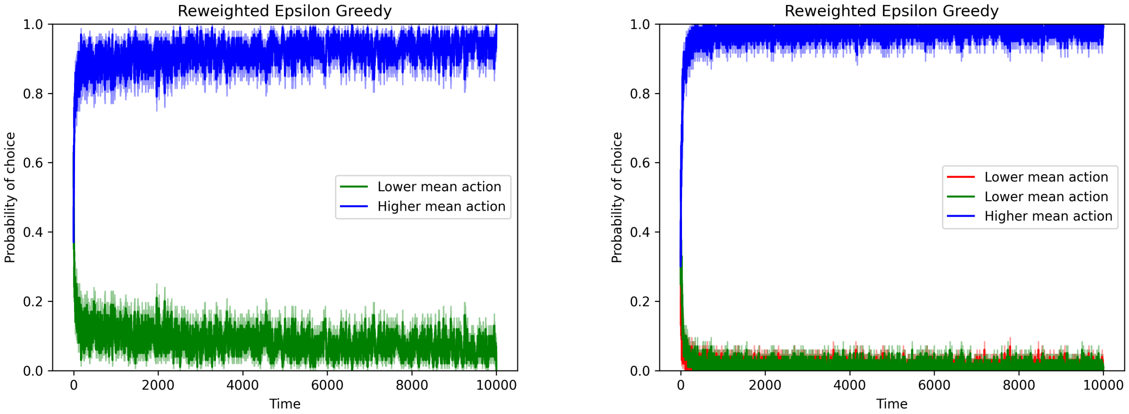
for some $t'' \in \mathbb{N}$ and $\tilde{t} \geq t''$. Choosing $t \geq \max\{t', t''\}$ yields the claim. □

Corollary 2.1. *ε-Greedy does not incur regret.*

Next, we provide a proof sketch that Optimistic ε-Greedy does not incur regret.

Proposition 2.1. *Optimistic ε-Greedy does not incur regret.*

Proof Sketch. This proof is similar to the proof that Upper Confidence Bound does not incur regret (see, *e.g.*, Auer (2003)). The only difference between the Upper Confidence Bound algorithm and the Optimistic ε-Greedy is exploration, which vanishes in the limit. It remains the case that the confidence bands are valid, and hence, there is a logarithmic upper bound for the probability that the bandit algorithm chooses a sub-optimal action in an exploitation step. As in the original proof of the Upper Confidence Band, this amounts to a sub-linearly growing probability of choosing a sub-optimal action and hence no regret. □



(a) No regret

(b) No regret

Figure 2.7: Plots illustrating the regret of Reweighted ε-Greedy. (a) For two arms, one of which with higher variance and one of which with higher expectation, the higher-expectation action is chosen most of the time. (b) For three actions with a deterministic action, a non-deterministic action of the same expectation, and a yet-higher-expectation dominant arm, we find that the dominant arm is chosen most of the time.

Finally, we test whether Reweighted ε-Greedy incurs regret. We present two simulations. One with an instance with payoff distribution Exp(1) and Exp(2) + $N(0,1)$, and another where two lower mean actions have payoff distributions $\mathbb{1}_{\{0.35\}}$ and $U[0.25, 0.75]$ while a higher mean action has payoff distribution $U[-1, 3]$. Figure 2.7 shows the results. We find that for a non-deterministic, non-centered dominant action, Reweighted ε-Greedy appears not to incur regret.

2.B Additional simulations

Both ε-Greedy's and Reweighted ε-Greedy's payoff estimates are biased estimators, compare Lattimore and Szepesvári (2020, Section 11.2). It is natural to ask whether debiasing the payoff estimates can address emergent risk preferences. To answer this, we run a simulation of a ε-Greedy with the statistic

$$\mu_{a,d}(t) = \frac{1}{N_a(t)} \sum_{\substack{1 \leq t' \leq t \\ a_t = a}} \frac{r_t}{\pi_{at}}.$$

This statistic leads to an unbiased estimate of the reward of an arm (Lattimore and Szepesvári 2020). The safe action in our simulation has reward distribution $\mathbb{1}_{\{0\}}$ while

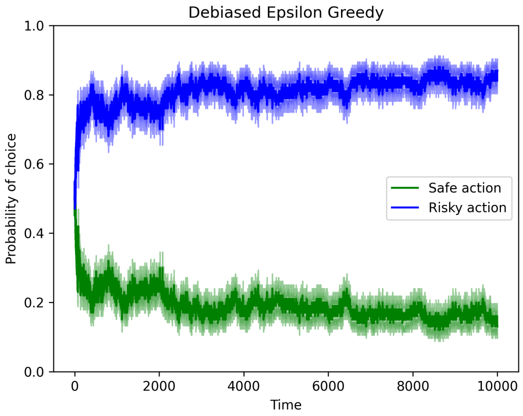


Figure 2.8: ε-Greedy with a debiased reward estimate. We find that for two centered actions ordered in second-order stochastic dominance, one of which is deterministic, the non-deterministic action is chosen most of the time.

the risky action has reward distribution $U[-1,1]$. Figure 2.8 reports the results. The debiasing leads to risk-*loving* behavior as opposed to risk-neutral behavior. One intuition for this uses monotonicity of risk attitude in probability normalization: The division by the square root of the choice probability in Reweighted ε-Greedy led to a correction that makes the algorithm exactly risk-neutral. Debiasing, which divides the statistic by the choice probability, a smaller number, leads to a stronger correction in the direction of risk-loving behavior.

Chapter 3

The Optimal Design of Generations

Chapter 2 examined how algorithms learn from repeated interactions with *different* users while learning to allocate optimally without prior knowledge of the type distribution F or the utility function u. This (brief) chapter discusses the optimal choice of personalized experiences for repeated interactions of the same user with the algorithm. In this chapter, *users* personalize by drawing a variable number of samples from a distribution designed by a designer, until the cost of taking an additional sample outweighs its expected benefit. We give an example where the optimal solution either does not show content preferred by a minority not at all, or shows it more frequently than minority incidence would suggest. Discussions of minorities and majorities are continued in the next Chapter 4.

3.1 Introduction

Users are not passive in personalization but take actions to shape the personalized experience through search. Consider, for example, a web search. A user, conditional on their search query, scrolls to the first entry that is sufficient for them to not warrant additional search. Different users have different preferences and will scroll for a different amount of time.

We consider a model of such search. Users are heterogenous and draw repeated samples from a distribution picked by a designer. What is the optimal distribution to generate outputs? It might be intuitive to output the majority's preferred content. We show an example where this is not the case.

Related Literature. Our model relates to consumer search. We consider a sequential search model in the tradition of Weitzman (1979). We also contribute to the literature trying to understand the role of randomness in optimal policies. Eysenbach and Levine (2019) discusses reasons why reinforcement learning algorithms sample actions randomly. One of their explanations, explained in their section 2 (the meta-partially observed Markov decision process) considers a similar setting to ours, but does not formalize user behavior as search.

3.2 Model

We model the search process of a user after they have given a prompt to a system. A user has unknown preferences represented by a type *type* $\theta \in \Theta$. The designer holds a prior $F \in \Delta(\Theta)$ over the user's type. The designer chooses a distribution $G \in \Delta(X)$ on an *allocation* or *generation space* X. The agent has a (von Neumann-Morgenstern) utility function $u \colon X \times \Theta \to \mathbb{R}_+$, which is common knowledge. We interpret 0 as a (normalized) utility of an outside option. Users incur a search cost of $c > 0$ per sample from G.[1] This search cost can be opportunity cost of time, or real cost, in case sampling from G is very costly (which is the case, for example, in generative models).

After the principal chooses G, the user repeatedly may choose to sample another draw and pay c, or stop and consume the best outcome they have seen so far, or the outside option, whichever is better.

Our main result assumes binary utility, which we also call *acceptable generations*. The reason for this name is that for binary utilities, types can be identified with the set $\theta \subset X$ of generations that give them utility 1. We call these *acceptable generations*. For binary utilities, we can write, as an abuse of notation $x \in \theta$ if and only if $u(x;\theta) = 1$, and call x *acceptable for* θ.

The designer wishes to maximize welfare, which is the final utility of the user, net of any search cost incurred.

A helpful quantity for the acceptable generations domain is $G(\theta)$, $\theta \in \Theta$, which is the mass of outcomes that are accepted by type $\theta \in \Theta$,

$$G(\theta) = \sum_{x \in \theta} g(x).$$

In the following, we give a tight characterization for a setting with no overlap in acceptable generations, *i.e.* for all $x \in X$, $x \in \theta$ for at most one $\theta \in \Theta$, which we call disjoint acceptable generations.

3.3 Optimal Generation

We now characterize optimal generations and show that they may represent minorities beyond their incidence in the population F. Observe that each agent with type θ searches until they reach a *reservation utility* $r(\theta)$ at which they are indifferent between continuing to search and getting utility $\mathbb{E}[\max(r(\theta), u(x;\theta))] - c$ and the utility of the best generation they have seen so far, $r(\theta)$. As Weitzman 1979 observed, it is optimal for agents to continue searching for any utility they have seen that is below $r(\theta)$ and stop searching for all higher utilities. This means that the designer's problem can be characterized as:

$$\sup_{G \in \Delta(X)} \mathbb{E}_{x,\theta}[u(x;\theta) | u(x;\theta) \geq c] - \mathbb{E}_\theta \left[\frac{1}{\mathbb{P}_x[u(x;\theta) \geq c]} \right]$$
$$\text{where } r(\theta) \text{ solves } \mathbb{E}[\max(r(\theta), u(x;\theta))] - c = r(\theta), \text{ for all } \theta \in \Theta. \quad (3.1)$$

[1] c may be correlated with θ, but we consider a fixed cost for all agents.

While this holds for general utilities, with binary utilities, we find the following.

Proposition 3.1. *Assume that utilities are acceptable generations. Then G is an optimal solution of (3.1) if*

$$\sup_{G\in\Delta(X)} \mathbb{P}[G(\theta) \geq c] - \mathbb{E}_\theta\left[\frac{1}{G(\theta)}; G(\theta) \geq c\right].$$

This objective clarifies a tradeoff in generation design: On the one hand, the designer would like to maximize the mass of agents that search and eventually get an acceptable outcome ($\mathbb{P}[G(\theta) \geq c]$) while minimizing the length of their search ($\mathbb{E}_\theta[\frac{c}{G(\theta)}; G(\theta) \geq c]$). If agents are excluded ($G(\theta) < c$), then it is too costly for other agents in terms of increased search costs to sacrifice probability to generate an acceptable outcome $x \in \theta$.

Proof. In the acceptable generations domain, there are only two types of search trajectories: either they search until they find an acceptable outcome, or they never start searching. The last group gets utility 0. For those that have not found an acceptable generation yet, there are the groups where the probability that an acceptable state $G(\theta)$ is drawn is higher than c, and these will search until they reach an outcome they accept, the other types will never draw a sample. This leads to a simplification of (3.1) as

$$\sup_{G\in\Delta(X)} \mathbb{E}_{x,\theta}[1; G(\theta) \geq c] - \mathbb{E}_\theta\left[\frac{1}{G(\theta)}; G(\theta) \geq c\right] = \sup_{G\in\Delta(X)} \mathbb{P}[G(\theta) \geq c] - \mathbb{E}_\theta\left[\frac{c}{G(\theta)}; G(\theta) \geq c\right]$$

□

The main result of this brief chapter is the following. It shows that minority agents who are not excluded are over-represented in optimal generations, in the sense that their probability is overweighted.

Theorem 3.1. *For any optimal solution G for generation design for disjoint acceptable utilities, for $\theta \in \Theta$, $G(\theta) \geq c$ we have*

$$G(\theta) = \frac{\sqrt{f(\theta)}}{\sum_{\theta: G(\theta)\geq c} \sqrt{f(\theta)}}.$$

For example, consider a setting where $\Theta = \{\theta_{\text{min}}, \theta_{\text{MAJ}}\}$ and $F(\{\theta_{\text{min}}\}) = 25\,\%$. In this case, for sufficiently low cost leading to θ_{min} being sampled at all, the optimal generation has $G(\{\theta_{\text{min}}\}) = \frac{1}{1+\sqrt{3}} \approx 37\,\%$, which is strictly larger than the minority incidence of 25 %.

This result means that the acceptable generations of statistical minorities will be sampled with higher probability than their incidence in the population if they are served. The statistical majority will be sampled less than their incidence in the population.

Proof. We restrict Θ to those types for which $G(\theta) \geq c$, which allows us to simplify (3.1) to

$$\inf_{G\in\Delta(X)} \sum_{\theta\in\Theta} \frac{f(\theta)}{G(\theta)}. \tag{3.2}$$

Note that as the types accept discount outcomes, we can reparameterize the problem from finding $G \in \Delta(X)$ to finding $G(\theta), \theta \in \Theta$ subject to $G(\theta) \geq 0$ and $\sum_{\theta \in \Theta} G(\theta) = 1$. The former constraints will never bind, as the objective diverges to infinity as $G(\theta)$ approaches 0 for any $\theta \in \Theta$. Hence, we may consider the Lagrange program for (3.2), whose critical points are given by

$$\lambda = -\frac{f(\theta)}{G^2(\theta)}, \quad \theta \in \Theta. \tag{3.3}$$

Here, λ is the dual variable for the normalization constraint. The solutions of (3.3) are on the ray $G(\theta) = \lambda\sqrt{f(\theta)}$, $\lambda \in \mathbb{R}$. Normalizing to a probability distribution yields the result. □

3.4 Discussion

The example provided in this brief chapter shows that user agency may impact concentration. In a binary-utility setting with disjoint generations that yield utility one, content for a minority is more frequently generated at optimality than their incidence in the population. Heterogenous search lengths drive the designer of generations to employ randomization and to increase probability with which minority-acceptable content is shown. We discuss assumptions and implications for settings where utility functions must be learned, before advancing to the next chapter, which will make an opposite prediction.

The main implication of this finding is that for personalization, additional understanding of user search processes will be important for the Economic Engineering of Personalized Experiences.

Different welfare weighting and independence of user prompting can be relaxed. Different welfare weights can be incorporated by rescaling the population distribution F. Similarly, it is possible to incorporate user prompting into the model. The generation design problems for different contexts are independent.

Two more assumptions are more substantial. First, we assume that users query the model repeatedly. This is an extreme assumption compared to one where a number of samples is requested and evaluated in batch, such as in (*Evaluating Large Language Models Trained on Code* 2021). It is unrealistic if the system has a good way to incorporate the search length of the user in its choice of generations. In such a model, a deterministic choice of personalized experiences will be optimal (as usual for partially observed Markov Decision Processes (Kaelbling, Littman, and Cassandra 1998)).

A final assumption is that model studied in this chapter assumes a user model (u, F). While this is a reasonable assumption in a recommendation system, it is currently not the case in many real-world (generative) systems, with a few exceptions, notably Kirk et al. (2024), Siththaranjan, Laidlaw, and Hadfield-Menell (2024), and Poddar et al. (2024).

We now continue with our next substantial chapter, which considers the role of preference measurement error for concentration, making a prediction counter to the example in this chapter: personalization increases concentration.

Chapter 4

The Distributional Consequences of Noise in Personalization

Chapter 2 studied the impacts of engineering choices regarding the personalization algorithm on the probability of showing a particular personalized experiences, the *algorithmic demand*. This chapter studies the case of optimal (one-shot) personalization. A user has a type $\theta \sim F \in \Delta(\Theta)$. An algorithm observes a noisy version of user preferences, s and chooses an allocation $x \in X$. We are interested in how preference measurement error in s affects the probability that statistical minority content is chosen and consumer welfare for minority users. We show that for symmetric measurement error, the recommendation is *lower* than the incidence of the minority, providing a different direction than our result from Chapter 3.

4.1 Introduction

Personalized experiences are ubiquitous in our everyday lives. From movie recommendations (*e.g.*, Netflix, Hulu) to short blogs (*e.g.*, TikTok, Twitter, Mastodon) and e-commerce (*e.g.*, Amazon), people turn to these recommendation systems to select entertainment, information, and products. For example, a recent study by Gomez-Uribe and Hunt (2016) revealed that 80 % of the approximately 160 million hours of video streamed on Netflix were recommended by one of the company's *personalization algorithms*.

In a classical recommendation system, a user's recommendations are based on signals on the preferences s, which can be based on user demographics, prior usage history, and explicitly given feedback. For simplicity, we can think of such signals as

$$s = \theta + \varepsilon(\theta),$$

where $\theta \in \mathbb{R}^d$ defines a user's preferences, and $\varepsilon(\theta)$ is a preference measurement error. A typical algorithm will select content $x \in X$ to maximize user utility $u(x; \theta)$. Hence, the algorithm tries to remove the impact of noise on recommendation.

The main result and theme of this chapter is that preference measurement, if symmetric across groups, aggravates concentration and inequality. In particular, in a two-group setting, content preferred by a statistical minority is allocated less by a Bayes-optimal algorithm compared to the minority's incidence in the user population.

Under general measurement error structures, the opposite may hold. Minority content may be recommended for any fraction from zero to twice the minority incidence. Minority welfare might be higher than majority welfare. The main assumption is that minority preference measurements may be significantly more accurate than preference measurements for the majority.

One of the main candidates for more informative signals to an algorithm is different signaling behaviors from users. We investigate such behaviors in a survey of 100 U.S.-based TikTok users on Amazon Mechanical Turk. We find that 30 % of users would change their behavior with the algorithm in an environment where they do not signal their preferences to the algorithm. We also show that users have a highly correlated understanding of how to signal to an algorithm. As an example, the following two users make decisions about some content to better signal their preferences to an algorithm.

> *"I make sure to interact [with] things that are specific to content types I want to see, even if I don't really love the content of that specific video."*

> *"If there was a private mode, I would use that to search things that I wouldn't want recommended to me. Stuff that I like, but stuff that I wouldn't want to clog my feed."*

The rest of this chapter is structured as follows. In Section 4.2, we review related work and the interpretation of noise terms in personalization. In Section 4.3, we present our model of recommendation in the presence of preference measurement error. Section 4.4 collects our theoretical results on concentration and inequity in personalized experiences. We present evidence from TikTok in Section 4.5. We discuss implications and extensions in Section 4.6.

4.2 Related Work

The first literature we relate to is on fairness in personalization. Popularity bias (Abdollahpouri 2019) is a statistical bias arising from not correcting for propensity in recommending content. The problem of recommending based on little data on the user is called the *cold start* and is solved with active exploration techniques (Safoury and Salah 2013; Zheng, Agnani, and Singh 2017). Some papers explicitly consider recommendations for statistical minorities, which are called *grey sheep users* (Alabdulrahman and Viktor 2021; Zheng, Agnani, and Singh 2017). Additionally, our work is related to users' efforts to improve their signal to personalization algorithms based on their understanding of the algorithm (Eslami et al. 2016; A. Y. Lee et al. 2022; Klug et al. 2021; Simpson, Hamann, and Semaan 2022). Previous papers Cen, Ilyas, and Madry (2024), Cen, Ilyas, Allen, et al. (2024), and A. Haupt, Podimata, and Hadfield-Menell (2023) considered the consequences of user exaggeration of their preferences. In contrast to these papers, we consider explicitly preference measurement error and market effects on concentration in personalization.

We also relate to simulation and empirical work in Industrial Organization. Close to this work, the simulation study Calvano et al. (2023) considers a two-sided market with a personalization algorithm. The paper's simulations feature significant measurement error—Calvano et al. (2023, Equation (5)) defines noise due to measurement error twice as large as their user heterogeneity. We showcase here how an optimal algorithm would favor some goods over others. More broadly, our study can be seen as contributing to a conversation on mass vs. niche content. Anderson (2006) argues that algorithms help a long tail, that is, very infrequently bought, items to rise to prominence. Fleder and Hosanagar (2009) takes the opposite perspective and points out additional concentration. We conclude that measurement error may or may not increase concentration, depending on how symmetric the preference measurement error is.

We also relate to equilibrium notions relying on noise added to preferences. Similarly, the notion of quantal response equilibrium (QRE), McKelvey and Palfrey (1995) relies on players observing a utility shock, optimizing based on it, but not observing other agent's utility shocks. Our model does not consider random utility shocks, but preference measurement error. Our model also relates to models of mechanism design with complex statistical types (Cai and Daskalakis 2022; Parkes, Ungar, and Foster 1998; Parkes and Ungar 2000). In contrast to these models, we do not allow the algorithm to decide on queries to the user, but we take measurement error as a primitive of the environment. Finally, our preference measurement error can be interpreted as a behavioral imperfection, and actions from users to improve signaling as sophisticated behavior, compare Laibson (1997), O'Donoghue and Rabin (1999), and O'Donoghue and Rabin (2001).

Finally, this work relates to algorithmic fairness. In particular, our comparison of the market share of minority content compared to the minority's incidence is mathematically equivalent to the recommendation algorithm's calibration gap (Pleiss et al. 2017). Our result on the incompatibility of fairness and efficiency, Theorem 4.1, can hence be interpreted as an instance of the incompatibility of accuracy and calibration. To achieve calibration in recommendation, Steck (2018) formalizes *item-level* calibration. Recommendations are item-level calibrated if the user sees items in a proportion that they consumed them in the past. Our results differ in that we consider the *population-level* distribution of recommendation.

Our techniques make use of (and our Propositions have corresponding results in) the literature on information design and Bayesian persuasion, compare Bergemann and Morris (2019). In fact, our model is mathematically equivalent to the classical Bayesian Persuasion model (Kamenica and Gentzkow 2011), but with a very different interpretation.[1] The assumption of symmetry we propose here has, to the best of our knowledge, not been studied in the information design literature.

4.3 Model

There are two types of agents $\theta_{\mathsf{min}}, \theta_{\text{MAJ}}$. A minority $\alpha < \frac{1}{2}$ is of type θ_{min}, a majority $1 - \alpha$ is of type θ_{MAJ}. We denote the set of types by Θ and probability distribution on types by

[1] The correspondence is the following: identify the majority type with the innocent state of the world, the minority type with the guilty state of the world, allocation of majority content with the acquittal action, allocation of minority content with a conviction action, measurement error with the investigation, the judge with the personalization algorithm, and the imagined adversary with the prosecutor.

$F \in \Delta(\Theta)$. A personalization algorithm chooses $x \in X := \{x_{\mathsf{min}}, x_{\textsc{maj}}\}$ for the users based on a noisy observation s of the type, distributed as $s \sim \sigma \colon \Theta \to \Delta(S)$. For our comparative statics in noise levels, we will use $S = \mathbb{R}$ and

$$\sigma(\theta_j) \sim N(\mu_j, \kappa^2).$$

for $j \in \{\mathsf{min}, \textsc{maj}\}$. The algorithm wishes to maximize user utility, which is binary:

$$u(x_j, \theta_{j'}) = \begin{cases} 1 & j = j' \\ 0 & \text{else.} \end{cases}$$

The timeline of the interaction is as follows. First, the user's type $\theta \sim F$ is realized. Next, the observation $s \sim \sigma(\theta)$ is realized and observed by the algorithm. Finally, the algorithm chooses $x \in X$ to maximize user utility.

4.4 The Consequences of Measurement Error

In this section, we investigate the probability of allocating minority content, $\mathbb{P}[x_{\mathsf{min}}]$, which we will call *minority share* or *algorithmic demand for minority content*, and the utilities of majority and minority users. A low probability of recommending minority content means a high amount of concentration.

4.4.1 Algorithmic Demand and Market Concentration

We first investigate how likely it is that minority content is allocated, $\mathbb{P}[x_{\mathsf{min}}]$. We compare it to the incidence of the minority in the population. For example, if a statistical minority is 10 % of the user population, we are interested in whether an optimal personalization algorithm will recommend content more or less than with 10 % probability. We first show that under general measurement error, it is possible that the minority anywhere from not at all to twice the minority incidence (*i.e.*, in our example, 20 %) is possible. It must be (weakly) less than the minority incidence for *symmetric* measurement error.

We first start with a property for general measurement error: It is possible that the minority share is anywhere from zero to twice the minority incidence. Any share in between is attained by some preference measurement error structure.

Proposition 4.1. *For any $p \in [0, 2\alpha]$, there is a measurement structure σ such that $\mathbb{P}[x_{min}] = p$. For any measurement structure, $\mathbb{P}[x_{min}] \leq 2\alpha$.*

The measurement errors that achieve the extreme cases of minority share are intuitive and we present them here. First, consider the case in which the measurement error is so large that it is pure noise. In this case, optimal personalization will serve majority content with probability 1 as it is more liked in the population. Hence, the minority share is zero under this measurement error.

The measurement errors that lead to higher-than-incidence minority share feature some asymmetry in their informativeness. While the minority incidence is, definitionally,

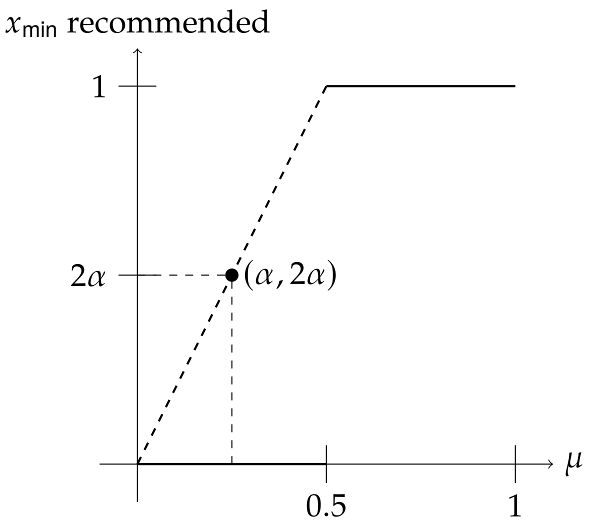


Figure 4.1: The information structure maximizing minority content allocation.

lower than the majority incidence in the population, it is possible that the minority incidence *conditional on a signal* is higher than the majority incidence. In this case, users with such a signal will be allocated minority content. The extreme case achieving 2α minority share is the case where conditional on all noisy observations of the minority type, the minority is in the majority conditional on the signal.

Proof. We use techniques from Bayesian persuasion (Kamenica 2019). We can view this problem as a setting where a sender chooses a preference measurement error σ. σ induces posterior probabilities over θ_{min} and θ_{MAJ}, which we can identify with a posterior distribution $\mathbb{P}[\theta_{\text{min}}|s]$ for $s \sim \sigma(\theta)$. Denote the distribution of these posteriors by μ_σ. We have that

$$\mathbb{P}[x_{\text{min}}] = \mathbb{E}_{x \sim \mu_\sigma}[\mathbb{1}_{x \geq \frac{1}{2}}]$$

We can use Kamenica and Gentzkow (2011, Proposition 1) to reduce this problem to choosing a posterior that, on average, is the prior, *i.e.* $\mathbb{E}[\mu_\sigma] = \alpha$. Such posteriors are also called *Bayes-plausible* posteriors in the literature following Kamenica and Gentzkow (2011). We can hence construct Bayes-plausible posteriors for the first part of the statement. To this end, we consider the Bayes-plausible posteriors that put mass p on $\frac{1}{2}$, and mass $1-p$ on $\frac{\alpha - \frac{p}{2}}{1-p}$. This is a Bayes-plausible distribution that achieves mass $\mathbb{E}_{x \sim \mu_\sigma}[\mathbb{1}_{x \geq \frac{1}{2}}] = p$. That is, probability p for classifying as θ_{min}.

For the second part of the statement, we use Kamenica and Gentzkow (2011, Corollary 1): The concave closure of the function $\mathbb{1}_{x \geq \frac{1}{2}}$, which is

$$\widehat{\mathbb{1}_{x \geq \frac{1}{2}}} = \begin{cases} 2x & x \in [0, \frac{1}{2}] \\ 1 & x \in (\frac{1}{2}, 1] \end{cases} \tag{4.1}$$

evaluated at α, hence 2α, see Figure 4.1. □

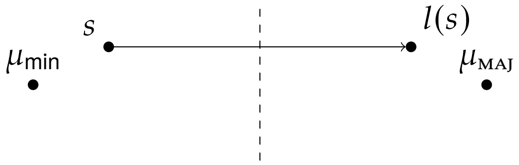


Figure 4.2: The midplane reflection.

In many environments, however, one does not expect such significant asymmetry, and we may have some structure where the observation probability of some signal conditional on types is the same. For this, we call a function $l\colon S \to S$ an *involution* if $l(l(s)) = s$ for all $s \in S$. The main example of an involution we consider is the *midplane reflection*. It is defined by

$$l(s) = \left(s - 2\frac{(s-\mu_{\text{min}})\cdot(\mu_{\text{MAJ}}-\mu_{\text{min}})}{\|\mu_{\text{MAJ}}-\mu_{\text{min}}\|^2}(\mu_{\text{MAJ}}-\mu_{\text{min}})\right),$$

and depicted in Figure 4.2. The existence of an involution means that there naturally are pairs $(s, l(s))$ in the signal space. (Note that $s = l(s)$ is possible and that the identity is an involution.)

Having an involution l, we can define *symmetric preference measurement error*.

Definition 4.1. We say that a measurement system $\sigma\colon \Theta \to \Delta(S)$ is *symmetric* if the density from the minority for a point s is the same as for the majority at $l(s)$,

$$\sigma(\theta_{\text{min}})(s) = \sigma(\theta_{\text{MAJ}})(l(s)). \tag{4.2}$$

By definition of an involution, a symmetric measurement error also satisfies $\sigma(\theta_{\text{min}})(l(s)) = \sigma(\theta_{\text{MAJ}})(s)$. A main example of symmetric measurement error for the midplane reflection l is $\sigma(\theta_j) = N(\mu_j, \kappa^2)$ for $\mu_j \in \mathbb{R}$ and some *common* variance κ^2 of the noise distributions.

Theorem 4.1. *Assume that a measurement system is symmetric. Then $\mathbb{P}[x_{min}] \leq \alpha$.*

Hence, under symmetric measurement error, the minority share is lower than the minority incidence.

Note that Theorem 4.1 holds for any involution l, in particular the identity function. For the identity function, $\mathbb{P}[\theta_{\text{MAJ}}|s] = 1 - \alpha > \alpha = \mathbb{P}[\theta_{\text{min}}|s]$ for any $s \in S$, and hence $\mathbb{P}[x_{\text{MAJ}}] = 1$ and $\mathbb{P}[x_{\text{min}}] = 0$. Hence, the minority share is zero in this case.

The intuition of the proof is to show that the signal pairs $(s, l(s))$ implied by the involution cannot both lead to the allocation of minority content. While this alone is not enough to conclude, we show that if one of them is a minority type, then the majority type must have a strictly higher likelihood.

Proof. Denote the set of signals $s \in S$ that are served x_{min} by S_{min}. That is,

$$S_{\text{min}} = \left\{s \in S : \mathbb{P}[\theta_{\text{min}}|s] \geq \frac{1}{2}\right\}. \tag{4.3}$$

Define $S_{\text{MAJ}} = S \setminus S_{\text{min}}$. We will write $\mathbb{P}[s]$ for the likelihood of s.

We first consider points $(s, l(s))$ and observe that at least one of the points must be in S_{MAJ}. Then, we show that if $s \in S_{\text{min}}$, then $l(s)$ has a higher likelihood than s. In the third step, we conclude.

Let $s \in S_{\text{min}}$. By Bayes' rule, it must be the case that

$$\frac{\alpha}{1-\alpha}\frac{\mathbb{P}[s|\theta_{\text{min}}]}{\mathbb{P}[s|\theta_{\text{MAJ}}]} = \frac{\alpha}{1-\alpha}\frac{\sigma(\theta_{\text{min}})(s)}{\sigma(\theta_{\text{MAJ}})(s)} \geq 1.$$

Hence, as $\frac{\alpha}{1-\alpha} < 1$, it must be that $\frac{\mathbb{P}[s|\theta_{\text{min}}]}{\mathbb{P}[s|\theta_{\text{MAJ}}]} > 1$. Hence, by symmetry, $\frac{\mathbb{P}[l(s)|\theta_{\text{MAJ}}]}{\mathbb{P}[l(s)|\theta_{\text{min}}]} > 1$. As $\frac{1-\alpha}{\alpha} > 1$,

$$\frac{1-\alpha}{\alpha}\frac{\mathbb{P}[l(s)|\theta_{\text{MAJ}}]}{\mathbb{P}[l(s)|\theta_{\text{min}}]} > 1,$$

and $l(s) \in S_{\text{MAJ}}$. Next we show, that if $s \in S_{\text{min}}$ and $l(s) \in S_{\text{MAJ}}$, then $\mathbb{P}[s] \leq \frac{\alpha}{1-\alpha}\mathbb{P}[l(s)]$. This follows from the following chain of inequalities:

$$\begin{aligned}
\mathbb{P}[s] &= \alpha\sigma(\theta_{\text{min}})(s) + (1-\alpha)\sigma(\theta_{\text{MAJ}})(s) \\
&\leq \alpha\sigma(\theta_{\text{min}})(s) + \alpha\sigma(\theta_{\text{min}})(s) \\
&= \alpha\sigma(\theta_{\text{MAJ}})(l(s)) + \alpha\sigma(\theta_{\text{MAJ}})(l(s)) \\
&\leq \frac{\alpha^2}{1-\alpha}\sigma(\theta_{\text{min}})(l(s)) + \alpha\sigma(\theta_{\text{MAJ}})(l(s)) \\
&= \frac{\alpha}{1-\alpha}\alpha\sigma(\theta_{\text{min}})(l(s)) + \frac{\alpha}{1-\alpha}(1-\alpha)\sigma(\theta_{\text{MAJ}})(l(s)) \\
&= \frac{\alpha}{1-\alpha}[\alpha\sigma(\theta_{\text{min}})(l(s)) + (1-\alpha)\sigma(\theta_{\text{MAJ}})(l(s))] \\
&= \frac{\alpha}{1-\alpha}\mathbb{P}[l(s)].
\end{aligned}$$

The two inequalities use that $s \in S_{\text{min}}$ and $l(s) \in S_{\text{MAJ}}$.

Hence, we can decompose $S = S_{\text{both}} \cup S_{\text{min}} \cup l(S_{\text{min}})$, where $S_{\text{both}} = \{s \in S | s, l(s) \in S_{\text{MAJ}}\}$. Note that this is a disjoint union. We hence have that

$$\mathbb{P}[S_{\text{min}}] \leq \frac{\alpha}{1-\alpha}\mathbb{P}[l(S_{\text{min}})] \leq \frac{\alpha}{1-\alpha}(1 - \mathbb{P}[S_{\text{min}}]).$$

Algebra shows that this inequality implies $\mathbb{P}[S_{\text{min}}] \leq \alpha$. As $\mathbb{P}[x_{\text{min}}] = \mathbb{P}[S_{\text{min}}]$, this concludes the proof. □

Hence, symmetric measurement error increases concentration. For a particular model of noise, we can show that not only is the minority share lower under measurement error than without measurement error but also that it is *decreasing* in measurement error.

Consider one-dimensional Gaussian measurement error

$$\sigma_\kappa(\theta_j) = N(\mu_j, \kappa^2).$$

It is without loss to normalize $\mu_{\text{min}} = 0$ and $\mu_{\text{MAJ}} = 1$. For such Gaussian measurement error, the minority share decreases as measurement error increases.

Proposition 4.2. *For not too large $\kappa < (\ln(\frac{\alpha}{1-\alpha}))^{-\frac{1}{2}}$, $\mathbb{P}_{\sigma_\kappa}[x_{min}]$ is monotonically non-increasing in κ.*

The condition on κ is mild. It holds for decision boundaries $x^* \leq -\frac{3}{2}$, far on the left of the minority type, meaning a rather extreme level of preference measurement error.

Proof. The decision boundary is given by an equality of likelihood:

$$\alpha \cdot \frac{1}{\kappa\sqrt{2\pi}} e^{-\frac{x^2}{2\kappa^2}} = (1-\alpha) \cdot \frac{1}{\kappa\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2\kappa^2}}.$$

Algebra yields that the decision boundary is

$$x^* = \frac{1}{2} + \kappa^2 \ln\left(\frac{\alpha}{1-\alpha}\right).$$

All signals $s \leq x^*$ are allocated x_{min}, all other signals are allocated s_{MAJ}. The probability that minority content is allocated is

$$\alpha\Phi\left(\frac{1}{2\kappa} + \frac{\kappa \ln\left(\frac{\alpha}{1-\alpha}\right)}{2}\right) + (1-\alpha)\Phi\left(-\frac{1}{2\kappa} + \frac{\kappa \ln\left(\frac{\alpha}{1-\alpha}\right)}{2}\right).$$

Here, Φ is the cumulative distribution function of a standard Gaussian. The first summand is the contribution of allocating for minority types, the second summand is for majority types. It is again a result of algebra that the derivative of this function with respect to κ is

$$\alpha\phi\left(\frac{1}{2\kappa} + \frac{\kappa \ln\left(\frac{\alpha}{1-\alpha}\right)}{2}\right)\left(-\frac{1}{2\kappa^2} + \frac{\ln\left(\frac{\alpha}{1-\alpha}\right)}{2}\right) + (1-\alpha)\phi\left(-\frac{1}{2\kappa} + \frac{\kappa \ln\left(\frac{\alpha}{1-\alpha}\right)}{2}\right)\left(\frac{1}{2\kappa^2} + \frac{\ln\left(\frac{\alpha}{1-\alpha}\right)}{2}\right),$$

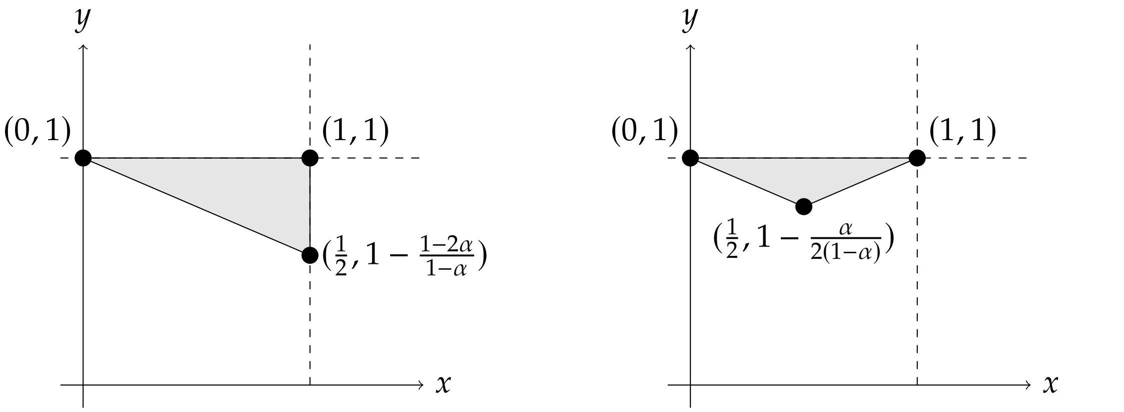
where ϕ is the probability density function of a standard Gaussian. As probability density functions are non-negative, and by assumption on κ, this function is negative, as required. □

This section contained three results: First, under general measurement errors, the minority share may be zero, smaller than the incidence of the minority, or up to twice as large. When measurement error is symmetric, minority content must be (weakly) less allocated than minority incidence in the population. Hence, minority content is disadvantaged compared to its incidence in the population. Finally, for (not too large) Gaussian measurement error, the minority share is decreasing in measurement error.

4.4.2 User Welfare

As in the previous subsection, we first show the user utilities possible under arbitrary measurement error structures.

Proposition 4.3. *The set of achievable user utilities for majority and minority under arbitrary measurement error is the triangle* $\operatorname{conv}(\{(0,1),(1,1),(1,\frac{1-2\alpha}{1-\alpha})\}$.



(a) General measurement error

(b) Symmetric measurement error

Figure 4.3: Achievable user utilities under measurement error.

Note that the triangle of achievable utilities for the minority and the majority is tilted. While minority utilities of zero (hence no minority agent is served their preferred content) are possible, there is a lower bound for the majority.

Proof. Note that by definition of the utilities, it must be that the set of achievable utilities is contained in $[0, 1]^2$. In addition, the set of achievable utilities is convex, compare Kamenica and Gentzkow (2011). As in the proof of Theorem 4.1, it is sufficient to (a) show that there are Bayes-plausible posterior distributions that yield utilities $(0, 1)$, $(1, 1)$, and $(1, \frac{1-2\alpha}{1-\alpha})$ and (b) show that $\frac{\alpha}{1-\alpha} u_{\mathsf{min}} + u_{\text{MAJ}} \geq 1$. (Note that $\frac{\alpha}{1-\alpha} u_{\mathsf{min}} + u_{\text{MAJ}} = 1$ is the line through $(0, 1)$ and $(1, \frac{1-2\alpha}{1-\alpha})$, compare Figure 4.3.)

To show (a), we observe that utility profile $(1, 1)$ is achieved by a Bayes-plausible posterior of

$$\mu_\sigma = \begin{cases} 0 & \text{w.p. } \alpha \\ 1 & \text{w.p. } 1 - \alpha, \end{cases}$$

which corresponds to no measurement error. Similarly, $(0, 1)$ is achieved by the deterministic posterior $\mu_\sigma = \alpha$, corresponding to "perfect" measurement error, that is, no information. Finally, $(1, 1 - \alpha)$ is achieved by the Bayes-plausible posterior

$$\mu_\sigma = \begin{cases} \frac{1}{2} & \text{w.p. } 2\alpha \\ 0 & 1 - 2\alpha. \end{cases}$$

When breaking ties at equal odds for both groups in favor of the majority, this leads to utility 1 for minority group users and utility $\frac{1-2\alpha}{1-\alpha}$ for majority users.

To show (b), it must be the case that for every agent who does not win from the majority group, there must be at least an equal mass of agents from the minority group that gets correctly allocated. Formally,

$$\begin{aligned} \alpha u_{\mathsf{min}} = \mathbb{P}[\theta_{\mathsf{min}}]\mathbb{P}[x_{\mathsf{min}} | \theta_{\mathsf{min}}] = \mathbb{P}[x_{\mathsf{min}}; \theta_{\mathsf{min}}] \\ \geq \mathbb{P}[x_{\mathsf{min}}; \theta_{\text{MAJ}}] = \mathbb{P}[\theta_{\text{MAJ}}]\mathbb{P}[x_{\mathsf{min}} | \theta_{\text{MAJ}}] = (1 - \alpha)(1 - u_{\text{MAJ}}). \end{aligned}$$

Rearranging, we obtain

$$\frac{\alpha}{1-\alpha} u_{\mathsf{min}} + u_{\text{MAJ}} \geq 1$$

as desired. □

For symmetric preference errors, we get a more restricted set of achievable utility profiles:

Theorem 4.2. *The set of achievable user utilities for majority and minority under symmetric measurement error is the triangle* $\operatorname{conv}(\{(0,1),(1,1),(\frac{1}{2}, 1-\frac{\alpha}{2(1-\alpha)})\}$.

Two observations of how symmetry leads to inequality are noteworthy: With symmetry it is impossible for the majority to have lower than perfect utility without leading to lower utility for the minority, and minority utility is decreasing faster than majority utility.

A second observation is about the dependence of the utilities on α. For α approaching 0, the triangle flattens, guaranteeing the majority a high utility.

The proof of this result uses a guess-and-check approach. We first use the bound on minority shares from Theorem 4.1 as an additional constraint, yielding a candidate set of achievable utilities, and then construct a symmetric measurement error that achieves this utility profile.

Proof. Observe that the set of achievable utility profiles is convex also for symmetric measurement errors. Indeed, by mixing the signal assignments, averages of utilities are possible.

Also, note that the utility profiles $(0,1)$ and $(1,1)$ can be achieved with symmetric measurement error (uninformative resp. perfectly informative). A symmetric measurement error structure achieving $(\frac{1}{2}, 1-\frac{\alpha}{2(1-\alpha)})$ uses three signals $s, l(s)$ and $\tilde{s}$ (where $l(\tilde{s}) = \tilde{s}$) and is defined as

$$\begin{aligned}
\sigma(\theta_{\mathsf{min}})(s) = \sigma(\theta_{\text{MAJ}})(l(s)) &= \frac{1}{2} \\
\sigma(\theta_{\text{MAJ}})(s) = \sigma(\theta_{\mathsf{min}})(l(s)) &= \frac{\alpha}{2(1-\alpha)} \\
\sigma(\theta_{\mathsf{min}})(\tilde{s}) = \sigma(\theta_{\mathsf{min}})(l(s)) &= 1 - \frac{1}{2} - \frac{\alpha}{2(1-\alpha)}.
\end{aligned}$$

In this case, for signal s, x_{min} is allocated, for $l(s)$ and $\tilde{s}$, x_{MAJ} is allocated. This leads to utilities $u_{\mathsf{min}} = \frac{1}{2}$ and $u_{\text{MAJ}} = 1 - \frac{\alpha}{2(1-\alpha)}$.

Algebra shows that the additional constraint of the triangle (that is, the left top side of the line through $(\frac{1}{2}, 1-\frac{\alpha}{2(1-\alpha)})$ and $(1,1)$ is given by $\alpha u_{\mathsf{min}} + (1-\alpha)(1-u_{\text{MAJ}}) \leq \alpha$. As $\alpha u_{\mathsf{min}} + (1-\alpha)(1-u_{\text{MAJ}}) = \mathbb{P}[\theta_{\mathsf{min}}]\mathbb{P}[x_{\mathsf{min}}|\theta_{\mathsf{min}}] + \mathbb{P}[\theta_{\text{MAJ}}]\mathbb{P}[x_{\mathsf{min}}|\theta_{\text{MAJ}}] = \mathbb{P}[x_{\mathsf{min}}]$, this follows from Theorem 4.1. □

Finally, we show that for a structured (Gaussian) measurement error, we get that inequality is increasing in measurement error. (This result is independent of the size of the error, in contrast to Proposition 4.2.)

Proposition 4.4. *$u(\theta_{min})$ is monotonically non-increasing in κ.*

Proof. As before, the decision boundary is given by

$$x^* = \frac{\kappa^2 \ln\left(\frac{\alpha}{1-\alpha}\right) + 1}{2}.$$

The probability that an agent of type θ_{min} is served x_{min} is given by

$$\Phi\left(\frac{1}{2\kappa} + \frac{\kappa \ln(\frac{\alpha}{1-\alpha})}{2}\right).$$

The derivative of this function with respect to κ is

$$\phi\left(\frac{1}{2\kappa} + \frac{\kappa \ln\left(\frac{\alpha}{1-\alpha}\right)}{2}\right)\left(-\frac{1}{2\kappa^2} + \frac{\ln(\frac{\alpha}{1-\alpha})}{2}\right).$$

As ϕ is a non-negative function and $\ln(\frac{\alpha}{1-\alpha}) < 0$, this is non-positive. □

In this section, we first observed that minority shares may be lower or higher than minority incidence but must be lower for symmetric measurement errors. Measurement error may lead to arbitrarily low minority welfare, but majority welfare is bounded from below. Symmetry tightens the lower bound on majority welfare and means that any measurement error that reduces utility for the majority needs to decrease the utility for the minority more.

We hence see a major role of symmetry of measurement error for inequity. Asymmetry may recover equity among user groups. The next section tests whether necessary conditions for asymmetric measurement error through *differential user behavior* are given.

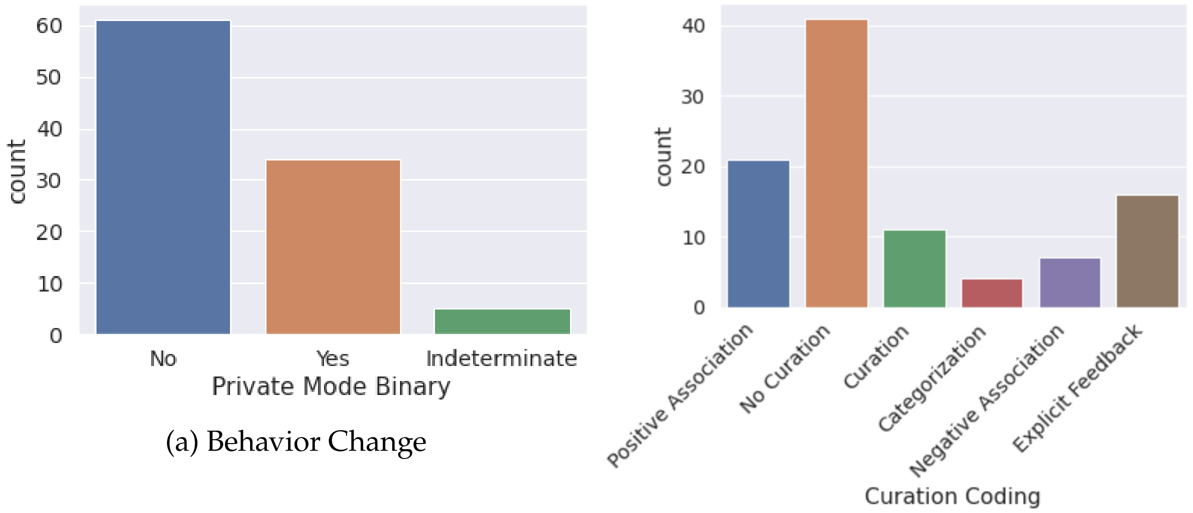
4.5 Evidence on Differential Signal Quality on TikTok

The last section highlighted the importance of measurement error asymmetry for concentration and user welfare. The first plausible source of asymmetry of measurement error is user actions to improve the outcome.

We test the necessary conditions for such actions to be taken: Awareness of measurement error and correlation between users when prompted to achieve a goal that allows for unambiguous communication of preferences.

Our survey was run on Amazon Mechanical Turk (short MTurk) during December 2022 and January 2023 (we reproduce the survey text in Section 4.D).[2] We solicited responses from 100 US-based participants regarding their interactions with content on TikTok. The participant pool is gender-balanced, mostly identifies as white, and mostly between the ages $30 - 59$ years old (see more information on the sample in Section 4.B). It is diverse in terms of content preferences, ranging from general categories like food and sports to mental health content and spiritualism. We list a full coding of content preferences in Section 4.A.2. We used pilot studies to manage experimenter demand concerns, see our methodology in Section 4.A.

[2] The data collection was deemed exempt under protocols E-4114 and E-4596 by the Committee on the Use of Humans as Experimental Subjects (COUHES) at the Massachusetts Institute of Technology.



(a) Behavior Change

(b) Types of Behavior Change

Figure 4.4: Changes in behavior under incognito mode

4.5.1 Incognito Mode

We first test awareness of behavior modification to influence measurement error. We ask respondents whether and how users change their behavior in the counterfactual of TikTok having a mode that does not allow to gather information on them during usage. We present the results in Figure 4.4a.

Around 30 % of the participants responded that they would change their behavior. To investigate whether this is a result of privacy concerns of users, or is related to signaling of their preferences, we analyze free-text responses. There are three roughly equally sized clusters among users that state that they would change their behavior. (i) A first group of users would engage more with content they currently do not want to be associated with (*e.g.*, “embarassing” or “risky” content), (ii) A second group states that they would engage more with content that they do not want to “clog” their feed, and (iii) a third group says that they would try to explore new content types that they have not been exposed to. We provide a breakdown of these counts in Table 4.1.

Incognito Mode Coding	**Participant Count**
no change: no reason	45
no change: less personalization	14
change: engage with “avoided” content	10
change: engage with “feed-clogging” content	9
change: exploration increase	9
other	8

Table 4.1: Encoding and statistics of the Incognito Mode question.

While (i) signals privacy concerns, (ii) and (iii) give clear signs of an awareness of preference measurement noise. The worry that their consumption of some content they like might lead to "feed clogging" is an instance of user expectation that the algorithm does not infer their preferences correctly from their behavior. Such feed clogging can be the result of an inference that users are part of a statistical majority liking such content where they aren't.

Similarly, (iii) points to users' concern about the noise of their preference measurement: exploration may lead to increased noise in the preference measurement.

Hence, there is evidence for awareness that there is measurement error.

4.5.2 Behaviors to Improve Signal Quality

Next, we ask participants in which ways they take actions to counter preference measurement error. Figure 4.4b shows the results. About 60 % of participants say that they *do* take actions to influence their future feed. 21 % state that they consume content they like to see more of (positive association), and 17 % say that they use explicit feedback mechanisms to the algorithm through likes, follows, etc. Other participants say that they consume content less that is similar to content that they would not like to see in the future ("negative association" in Figure 4.4b) or engage in both positive and negative association ("curation" in Figure 4.4b). We call "categorization" types of behavior in which users state that they consume content that they do not like because of worries that the algorithm associates them not liking this content with preferences over content that they do not like.

This means that a majority of users take actions that improve the measurement error of their preference consumption.

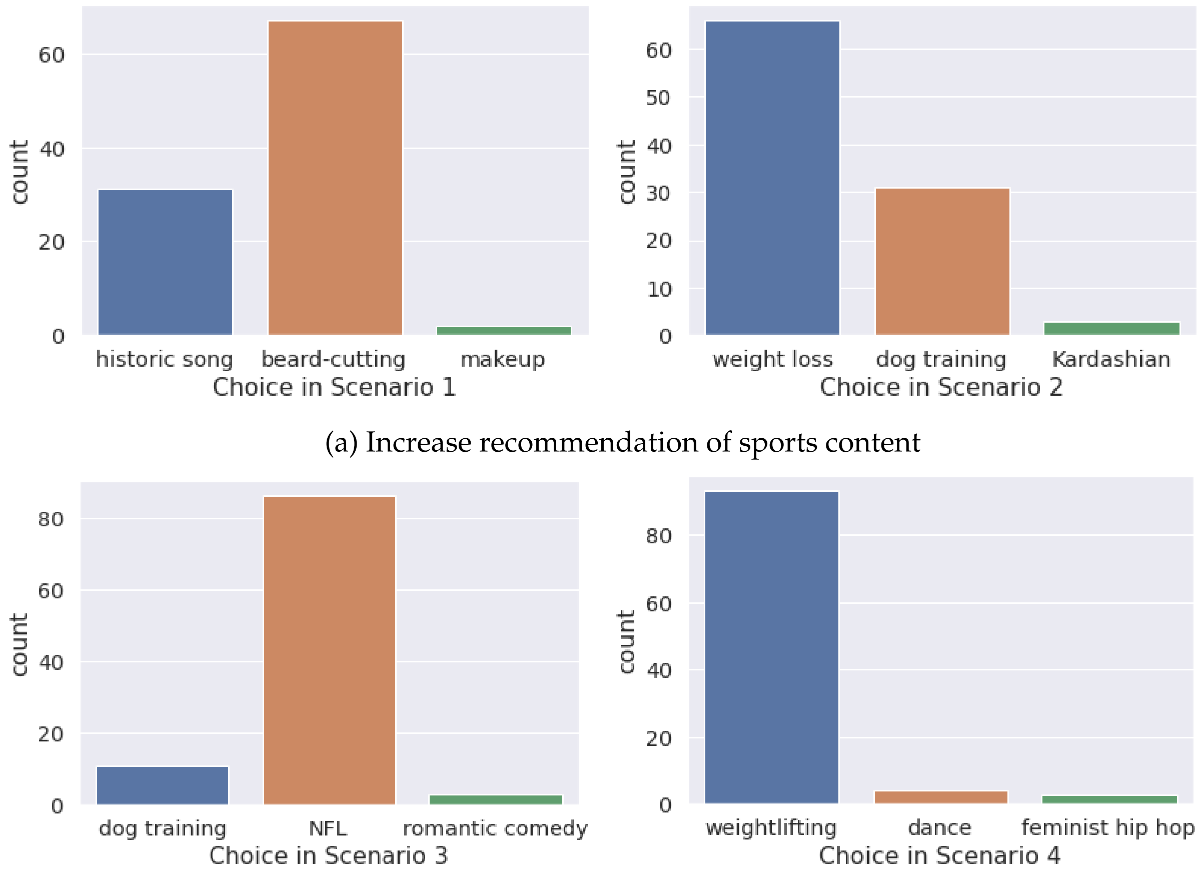
4.5.3 Hypotheticals

In the last part of the survey, we elicit whether users are capable of signaling their preferences with low noise by testing whether responses to some scenarios are correlated among users. We do so by presenting users with four different types of content whose recommendation frequency they are tasked to increase. We find significant correlation in answers, see Figure 4.5.

In the first and the second hypothetical, we asked the participants to choose the content they would consume in order to increase the sports content that the platform was recommending to them. We report results in Figure 4.5a. In both cases, the majority of the participants associated BEARD-CUTTING and WEIGHT LOSS with sports.

In the third and fourth hypothetical, we asked the participants to choose the content they would consume if they wanted to see less make-up content. We report these results in Figure 4.5b. In both scenarios, the majority of the participants chose NFL PLAYS and WEIGHTLIFTING as different from MAKE-UP.

Observe that the correlation between answers is higher in the second set of hypotheticals than in the first set. Free-text answers suggest that this is due to gender associations of makeup content, see Figure 4.6. Other responses, like other stereotypes, point in a similar direction.



(a) Increase recommendation of sports content

(b) Decrease recommendation of makeup content

Figure 4.5: Participants' aggregate responses to four scenarios for strategies to increase content they are served with in the future. The detailed scenarios can be found in Section 4.D.

Hence, we find that not only are users aware, but also, in principle, capable of providing a better signal to the algorithm.

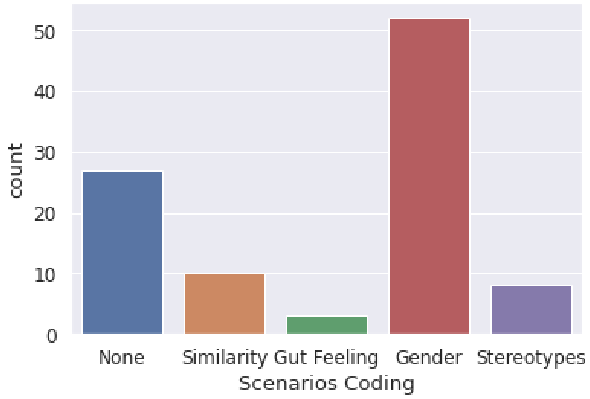


Figure 4.6: Aggregate responses to how the participants made the associations in the "Scenarios" questions.

4.5.4 Social Desirability and Experimenter Demands

We conclude this section by discussing two biases, social desirability, and experimenter demands, that might limit the evidence we can derive from our survey.

Social desirability bias means that users respond in a way that is socially desirable but not according to their own authentic interests. It is a main concern for our question on the incognito mode. It is possible that we underestimate the fraction of users that would change their behavior because of privacy concerns in a version of TikTok that does not store information persistently. Our main interest, however, is in those who would change their behavior to avoid feed clogging or increase their exploration. It is unclear whether social desirability affects the statement of such preferences. There are no apparent social desirability concerns in our questions on behaviors to improve measurement error and the hypotheticals.

Experimenter demand bias is the bias arising from participants responding in a way they predict the researcher would "like" as opposed to according to their own authentic interests. Our survey design, outlined in Section 4.A, tried to limit experimenter demand concerns. In the question on differential behavior in an incognito TikTok, user may have overstated whether they change their behavior on such a platform as this is in the interest of the experimenters. Similarly, users might overstate the actions that they take to reduce preference measurement error due to experimenter demand bias. There are no apparent experimenter demands for the hypotheticals.

4.6 Discussions and Avenues for Future Work

We studied the fairness implications of preference measurement error in personalization. We show that concentration is not generally higher due to measurement error, but it must be if errors are symmetric. We also show that under general measurement errors, minority welfare may be higher than majority welfare, but that this fails to be possible under symmetric measurement error.

We made several simplifications in our analysis. On a real platform recommending content to users, there are conflicts of interest between the platform and the user, there is no perfect conflict between the platform and the user, the type distribution needs to be inferred from data, and both users and content creators make strategic choices which we do not model. We comment on each of these in the following subsections and mention avenues for future work.

4.6.1 Conflict of Interest

We assumed that the personalization algorithm would like to maximize instantaneous user utility. Our results might change when taking into account that platforms would like to retain minority users. Measurement error that leads to minorities not getting their preferred content with a very high probability incentivizes them to leave the platform. We leave a model with such participation decisions for future work.

Even with a conflict of interest, a main result remains if we replace user utility with a platform objective: Symmetric measurement error will lead to increasing the objective that the platform has for the majority compared to the minority.

4.6.2 Popular Content

We considered a setting where there are two types of content and no content that is liked by both groups. This is for analytical simplicity, and our results in more than binary settings should be interpreted locally: In areas with noisy and symmetric preference areas, the content marginally liked by a statistical minority will be recommended less. Minority users may more often get less personalized content because of symmetric preference measurement error structures.

4.6.3 Utility Inference

We assumed in our model that the platform knows the utility function u and the prior F. When estimation is based on a linear regression, it is known that typical estimators are unbiased even in the presence of asymmetric noise, that is, heteroskedasticity (Angrist and Pischke 2008, Chapter 3). In nonlinear models such as those in industrial organization (*e.g.*, the Berry inversion, see S. T. Berry (1994)) or language model fine-tuning (Rafailov et al. 2023), this may not hold, and models of measurement noise are necessary for unbiased estimates.

4.6.4 Strategic Choices of Users and Content Providers

A final and large area that we do not consider is that actions are chosen strategically by users, but also content providers.

Demand Side

On the demand side, it may be that users not only choose to reduce the noise in preference estimation but even choose to exaggerate their consumption behavior to provide a stronger (and not only less noisy) signal to the personalization algorithm. Such behavior would lead to *stereotyping* of users. The empirical significance of such behaviors is an open problem which we leave for future work. As a microeconomic modeling question, it relates to the literature on strategic communication with lying costs (Kartik 2009; Deneckere and Severinov 2022), noisy signaling (Blume, Board, and Kawamura 2007; Landeras and Villarreal 2005; de Haan, Offerman, and Sloof 2011; Ying Chen 2011) and learning with strategic data sources (Hardt et al. 2016).[3]

Supply Side

A final unmodeled aspect comes from the role of content creators who supply content. Two participants refer to the role of content creators:

> *"I imagine there just aren't very many people in this field who are also inclined to make TikTok videos, or at least videos about their profession."* [sic]
>
> *"I'm not sure if TikTok is not showing me this type of content. I'm starting to believe there just aren't TikTok creators who fall into this category? I see representations of all kinds of people who identify as LGBT, I just very, very rarely see anyone my age who is super feminine like me. There [are] tons of feminine gays on TikTok in their mid-20s and below, just not in my age range."*

Our results on the minority share shed light on why there may be less minority content. As earnings from content often come in the form of revenue sharing based on the number of views of content, the lower minority shares under symmetric measurement error directly lead to less (monetary) reward for content creators working for niches in settings where preferences are hard to measure. This leads to efforts for content creators to make it easier for their content to be targeted.

4.A Methodology

In this section, we document our survey methodology.

[3] Other papers in the literature on learning with strategic data sources are the following: Hardt et al. (2016), Dong et al. (2018), Yiling Chen, Y. Liu, and Podimata (2020), Ahmadi et al. (2021), Ghalme et al. (2021), and Levanon and Rosenfeld (2021) bound the loss in objective from strategic choices of data, Meir, Procaccia, and Rosenschein (2012), Cummings, Ioannidis, and Ligett (2015), Yiling Chen, Podimata, et al. (2018), Ball (2025), and Eliaz and Spiegler (2019) provide guarantees on incentive compatibility of inference, and Hu, Immorlica, and Vaughan (2019), Milli, Miller, et al. (2019), and Bechavod et al. (2022) study fairness in (noiseless) models of learning with strategic data sources. Braverman and S. Garg (2020) shows that noise in the choice of the classification algorithm can help both the efficiency and fairness of the outcome. In contrast, we consider the role of (undesigned) measurement error of probabilities.

1. In the first phase, we brainstormed different categories of questions (concept-driven approach). Our goal was to find the "correct" set of questions that would simultaneously achieve two goals. The first goal was to not be too leading (*e.g.*, we did not want to ask explicitly whether they strategize with their content consumption to lead the algorithm to form specific associations with the content it was suggesting to them). The second goal was for our survey participants to understand the types of behaviors that we were asking them about (for example, we never used the word "strategize" in our survey). After every brainstormed set of questions, we ran a *pilot study* of 10 participants. Based on the responses we got each time, we calibrated our questions (and the categories of questions more broadly) until we converged to the ones we present on Section 4.D (data-driven approach).

2. Once we had converged to the set of questions, we deployed the survey on MTurk and solicited 100 full responses. While the survey was running, we made no changes to the set of questions. We chose to deploy our survey on MTurk since the biases and the demographics of the population of workers have been well-documented in the literature (see *e.g.*, P. G. Ipeirotis 2010; Hitlin 2016; Difallah, Filatova, and P. Ipeirotis 2018 for case studies).

3. In the third phase, we did qualitative data analysis using standard methods (see *e.g.*, Kuckartz 2019), which we outline next. All quantitative data (*i.e.*, demographics) is reported with no preprocessing. For the responses that were in free-form text, we inductively created specific *codes* that represented the common points in the participants' responses but were abstract enough so that they could include multiple responses. This coding step was required in order to obtain aggregate statistics from free-form text responses to our survey. The codes and their explanations for the different categories of questions can be found in Section 4.A.2.

4.A.1 Survey Design

One of our first steps in building the survey was deciding how the questions would be organized in blocks with shared goals.

The first block of questions addresses the participants' usage of the platform and the time they have spent on the platform since its adoption. The purpose of this block was to assess whether our respondents spend enough time on the platform so as to have started building folk theories about how the algorithm categorizes them and decides which content to serve to them.

The second block of questions asks the users about the types of content that they usually see on TikTok. We asked the participants both for the categories of content that TikTok puts more frequently into their feed and the specific subcategories in which they were mostly interested. Our goal here was to make the participants start thinking about the positive associations that the algorithm may be building with the types of content that they are interested in and the types of content that it puts into their feed.

The third block of questions asks users to report the parts of their identity that were not well represented by the types of content that TikTok suggested. Then, we solicited free-form text responses regarding their best explanation for why this happened. Our goal here was to have the participants start thinking about whether *they* take any actions to curate their feed that may have resulted in the algorithm presenting to them the type of content it currently does.

In our fourth block, we asked them whether a "private mode" on TikTok would make them change anything in the way that they interact with the platform. We again solicited free-form text responses. Our goal here is to understand whether participants are consciously stopping themselves from interacting with particular types of content out of fear that the algorithm will make associations that they do not want it to.

The final block of questions explicitly asks whether the participants take any actions to "curate" the type of content they see and tests whether participants understand how the algorithm makes associations and categorizes people based on the content they consume. The goal here was to give them one more chance to think about their own curation efforts, especially while being explicitly prompted to address these questions. The questions about associations were to give the participants specific examples of how the algorithm may pattern match between topics so that they could address it in the following question (*i.e.*, the explanation of why they thought that the algorithm would associate these topics).

4.A.2 Qualitative Analysis

In the following, we explain our codes for free-form text responses from the survey participants to questions (19), (20), and (25) (see Survey questions in Appendix 4.D).

Codes for Private Mode Question

No change for no stated reason Users would not change their behavior as they see no obvious reason to do so.

No change for personalization reasons Users would not change their behavior so as not to change their personalization of the algorithm.

Change to engage with "avoided" content Users would interact with content they don't want to be associated with normally (*e.g.*, embarrassing, risky content).

Change to engage with "feed-clogging" content Users would interact with the content they don't want to clog their feed.

Change to increase exploration Users would increase their exploration of new topics.

Other Users would change their behavior in other ways; for example, they would switch to other platforms (because their experience would worsen as a result of less personalization), or they would engage more with content they already like.

Codes for Curation Question

The following are the codes we derived for the question on a private mode:

Positive Association Users either watch more or for longer, videos that are similar to the types of content that they would like to see.

Negative Association Users either watch less or shorter videos that are unlike the types of content that they would like to see.

Curation Users describe that they engage in behaviors using both negative and positive associations.

Categorization Users specifically give evidence that they consume content that they do not like to not be categorized as not liking this type of content.

Explicit Feedback Users describe explicit ways to give feedback to the algorithm: Likes, follows, "I am not interested" buttons, blocking.

No Curation Users state that they do not curate content.

Codes for the Reflection on Scenario Questions

Gender Users state concretely that they would choose gendered content.

Stereotypes Users state that there are stereotypes and associations. For example, a historic venue in the music scenario might also be used for physical activity, hence connecting to sports.

Similarity Users state that they choose content that is similar or opposite, referring to the closeness of content.

Gut Feeling Users state that their choice was intuitive.

None Users reiterate their choices while not giving reasons for them, give evidence that they are not responding under the stated hypothetical preferences but their own, or in other ways do not give a clear reason for their choice.

4.B Sample Description

In this section, we report our survey participants' statistics: demographics, use of the platform, and topics they are interested in.

4.B.1 Demographics and Usage

The demographics of the survey participants are shown in Figure 4.7. Most of the participants (more than 80 % are white (which is in line with the population breakdown of MTurkers Difallah, Filatova, and P. Ipeirotis 2018). Around 50 % of the participants are *between* 40 *and* 69 *years old* and more than 40 % are *between* 30 *and* 39 *years old*. More than 50 % of the participants self-identify as *women*, and most of the rest self-identify as *men*. We also have a small representation of folks who self-identify as *genderqueer* and *non-binary*.

In terms of educational level and occupation, the majority of our survey participants have obtained a *Bachelor's degree* and are currently *employed for wages*.

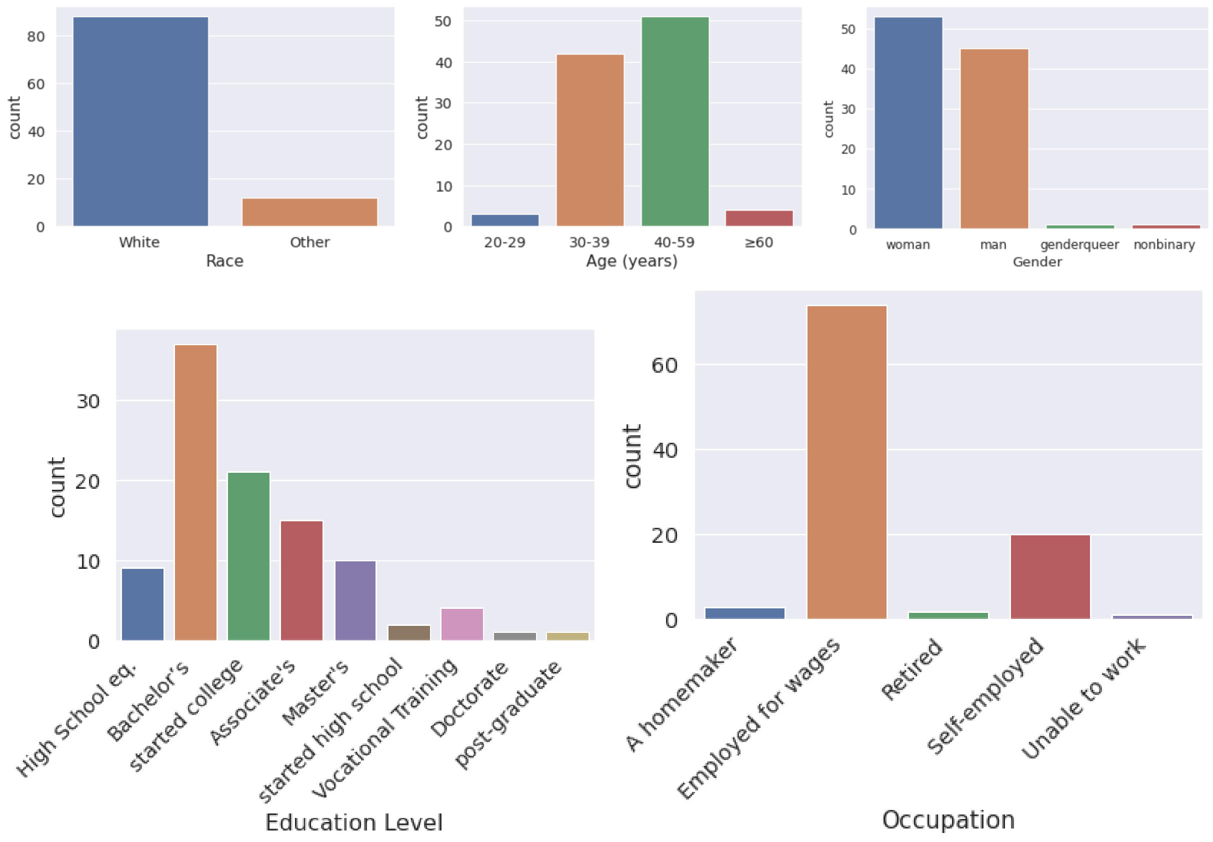


Figure 4.7: Demographics of survey participants. The x axis corresponds to the categories for each plot and the y axis reports the number of participants per category.

The usage statistics for the survey participants are shown in Figure 4.8. We see that the majority (more than 70 % have been using the platform for *more than a year*. Around 50 % of the participants use the platform daily, and the vast majority of all of our participants use the platform for less than 2 hours daily.

4.B.2 Content Consumption

In Table 4.2, we present the different primary content types that TikTok puts on the survey participants' feeds. We coded 34 distinct primary types of content that TikTok suggests to users from a broad set of content types.

4.C Selected Quotes

This section contains additional quotes from survey participants.

Content Types	**Participant Count**
food	48
funny	46
animals	25
sports	20
hobbies	17
family	15
dance	12
DIY	9
politics	9
music	9
film & TV	9
fashion	8
home improvement	7
money	7
exercise	6
mental health	6
science & technology	6
spiritual & religion	5
gaming	5
news	5
beauty	5
pop culture	3
health	4
work	2
history	2
LGBTQ+	2
educational	2
films & TV	1
travel	1
books	1
spiritualism & religion	1
conspiracy theories	1
art	1

Table 4.2: Content preferences in sample

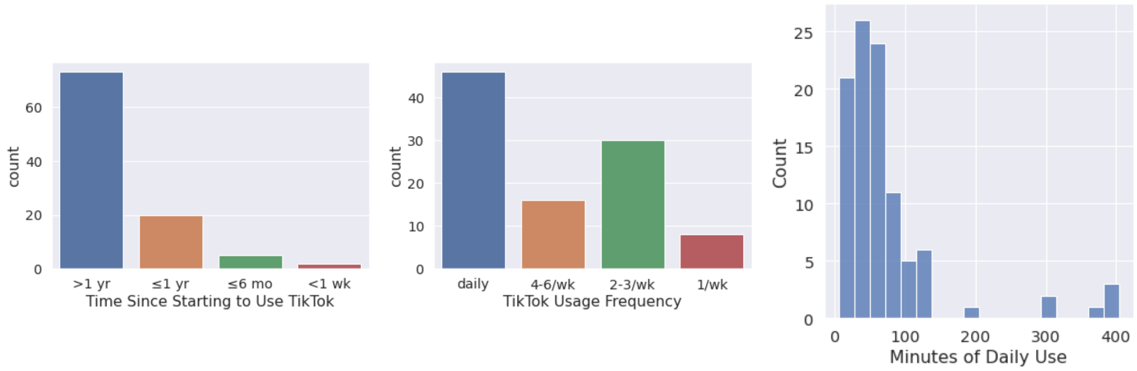


Figure 4.8: TikTok usage statistics of survey participants. The x axis corresponds to the categories for each plot and the y axis reports the number of participants per category.

4.C.1 Quotes on Representation of Identity and Curation

> *"I feel that Tiktok continues to put these in my feed because I almost always get sucked into watching them. That tells the algorithm that I like them, even though I am mostly just using them for background noise and have seen most of them before."*
>
> *"I follow a couple of people with barns, particularly horse rescues. I often like posts not just from them but other related content that I see"*
>
> *"I view this content sometimes unintentionally, but algorithmically it's recommended based on this + stuff like demographics, trends, area, etc."*
>
> *"I typically use TikTok as a way to relax and unwind, so I watch a lot of humorous content. I think TikTok uses my watch history to fill my feed with similar videos, like the bloopers I watch regularly."*
>
> *"I was looking into finding the right form for workouts. Now tiktok probably thinks I love powerlifting."*
>
> *"TikTok puts items related to movies into my feed because they track my viewing history and also videos that I comment on and like. There is an algorithm that runs in the background and collects this information and then sends me more of the same."*
>
> *"I think I have seen less of this on TikTok for two reasons. The first reason is that there is less clean comedy on TikTok. Reason number two is that it is often hard to tell when scrolling which videos will have clean stand-up. So I end up watching lots of stand-up, which does a poor job of training the TikTok algorithm."*
>
> *"I believe its because I already follow or view many of these type of TikToks so my feed is constantly showing me that type of content"*
>
> *"TikTok probably puts this into my feed because it is close to the eating challenges videos. I am interested in that, so it probably ties the two together. Also, I will watch those videos."*
>
> *"I look at this information daily on other social media. I suspect my information is being sold. I also view it on TikTok."*

"I think because I liked a video once of this type of content. I believe by me liking the video, the algorithm thought I would like to see more videos like that one."

"I have watched/viewed this content, so it makes sense from an algorithm-based standpoint. I also believe they will randomly push videos just to have a more broad focus on varying content types for all."

"Tikok believes that I really enjoy jump scare videos because I may have watched some in the past, either on TikTok or on YouTube. TikTok probably sees what I have watched on all of my platforms before and curates its suggestions based on that. If there is some kind of pattern it sees, it will seize on that and show you a bunch of videos in that genre."

"I'm constantly looking at video game TikToks, and some of those happen to be speed run related. It's easy for the algorithm to suggest that combination for my feed."

"I interacted a lot with videos about the racial justice protests in 2020. And then I would get more academic videos about anti-carceral theory as well."

"Because I have interacted with this content before. I have either liked this type of video or commented on this type of video."

"I respond well and engage with funny videos. The AI learns from that."

"I am always seeking out recipes. On most cooking videos I like and comment on the videos, so I stay on that side of TikTok."

"They put this into my feed because the algorithm picks up on when I stop and watch long-form sports videos on their platform. It continues to show me this information because it knows that I like it."

"I was probably annoyed by a specific creator who I always skipped, and now it thinks I don't want any videos on the topic."

"I have clicked not interested in this content. I also do not interact with it or watch it all the way through."

"Perhaps I haven't trained the algorithm enough to let it know it is something I am passionate about. They probably are trying to get me to watch other things related to stuff I have watched previously."

"I don't really view the internet very often for this specific lifestyle choice."

"I just do not report whether I am single or married."

"I don't know. I actually have an acute interest in religion, but I prefer my religious engagement to be somewhere less frivolous than TikTok."

"Again, I don't look up videos about retail because I'm at work most of the time and don't want to see work-related stuff in my free time."

"I don't follow any TikTok hikers or anything like that. TikTok doesn't know I have an interest in it."

"I don't watch a lot of videos or consume a bunch of content relating to exercise on social media platforms, so TikTok won't have that base to figure out that that's something I like. I enjoy exercise, but I do it on my own and don't post much about it."

"I suspect because I rarely search for any travel items on TikTok, I tend to do most of that on YouTube."

"If I like a certain type of video, I will make sure to like it. I will also leave a comment. Sometimes I will watch it twice, so the algorithm realized I am into that kind of content."

"I try to scroll past a video quickly if I don't like it because I don't want the algorithm to think I like this kind of content."

"I report offensive content and try to like things I enjoy. I'm also careful which things I share."

"I talk about topics in front of my phone, google stuff I want in feed. I also like stuff just to see more of that type of stuff even though I don't like it. Like sometimes, if my content gets too dark, I try to like animal videos and comedy more to get off the darker content for a bit."

"Currently, I am cognizant of what category of video I think material falls under. I am careful to watch complete videos that fall under the correct category (even if I am not interested in that particular video). I am careful to skip over videos from the 'wrong' categories. And I make sure to close down the app on a 'wrong' category."

"I would try to use my own imagination regularly to fine-tune my feed for this media site."

"Sounds odd, but sometimes I will click on something once, get out and click it again, then search for it if it's a topic I like that I realize I haven't been seeing. I don't know if that actually works, but it seems to."

"I am intentional with what I search. I want stuff that serves my purpose to me to try and not clog up my feed with stuff that I am genuinely not interested in."

"I try not to watch a video all the way through if it is something I am not interested in. Also, I try to swipe quickly on that content so as not to set off the algorithm."

"I make sure to interact with things that are specific to content types I want to see, even if I don't really love the content of that specific video."

4.C.2 Quotes Regarding Associations

"I feel like I need to trick the system into what I do and don't want to see. I feel like if I want to see less of something, I need to change what I look at to make them realize I don't want to see certain things. Since I don't wear makeup, I should be looking at more masculine things to stop from seeing makeup, for instance."

"There are subgroups that cluster together online according to gender association. Stereotypically feminine pursuits such as makeup do not tend to occupy the same area as sports. However, I wasn't as certain of the first, beard-cutting seems awfully sedentary for someone who wants to see sports, even if it is only of interest to males."

"I tried to think about how to 'trick' the algorithm into going in another direction. I think choosing the options that would get me there sooner is smarter. I think a too drastic change, however, might not stick with the algorithm" [sic]

4.D Survey

The survey consists of six question blocks:

1. The first block collects information regarding the respondent's platform usage. Specifically, we ask them to choose from a list of pre-defined responses how often they have used TikTok in the last 30 days and when they started using the platform. For days they did use the platform, we also asked them to report how long they did.

2. The second block consists of questions about the types of content that TikTok puts on the respondent's feed *most frequently*. Specifically, we first solicit answers on the top 3 types of content that the platform suggests, and then we ask for each of these content types which specific subtype *they* are more interested in.

3. The third block focuses on the accuracy of the model that TikTok has built for the respondents. We make this association of the algorithm's accuracy through the content types that are more/less frequently put on the users' feeds. Specifically, we first solicit responses regarding parts of the respondents' identities that have *not* been well-represented by TikTok's algorithm. Subsequently, we ask them to give their best explanation in free-form text about why the algorithm suggests their top 3 content types and why it is failing to represent well the parts of their identity that they declared at the beginning of the current block of questions.

4. The fourth block of questions asks whether the respondents would change anything in their TikTok usage if TikTok offered a "private mode". We also solicit free-form text responses with short explanations regarding their choice.

5. The fifth block of questions focuses on the content curation from the side of the users. We first ask them to free-form text responses regarding whether they consciously take actions to curate the content they are seeing on TikTok. Subsequently, we ask them to choose from a list of predefined answers what types of content they would consume more/less of if they wanted to signal to the algorithm that they want to see more/less of a particular type of content that is related.

6. The last block of questions elicited sociodemographic characteristics, including ethnicity, education level, professional status, age, and gender.

4.D.1 General Consumption

1. How often have you used TikTok in the last 30 days?

   (a) daily

   (b) 4-6 times a week

   (c) once a week

   (d) less than once a week

2. On days you use TikTok, how many minutes do you use it?

3. When did you start using TikTok?

   (a) less than a week ago
   (b) less than a month ago
   (c) less than six months ago
   (d) less than a year ago
   (e) more than a year ago

4.D.2 Content Consumption

The following questions will be about content, *e.g.*, videos, clips, photos, memes, or text, that you watch, look at or read on TikTok.

4. Which types of content does TikTok put into your feed **very frequently**? (Examples: Sports, Fashion, Food, History.) Please give three examples.

5. Which types of content does TikTok put into your feed **very frequently**? (Examples: Sports, Fashion, Food, History.) Please give three examples.

6. Which types of content does TikTok put into your feed **very frequently**? (Examples: Sports, Fashion, Food, History.) Please give three examples.

7. Consider the content of the type of your first example. Which more specific sub-type of content are you particularly interested in? (Examples within "sports": workout, baseball, ballet)

8. Consider the content of the type of your second example. Which more specific sub-type of content are you particularly interested in? (Examples within "sports": workout, baseball, ballet)

9. Consider the content of the type of your third example. Which more specific sub-type of content are you particularly interested in? (Examples within "sports": workout, baseball, ballet)

4.D.3 Fidelity of the Recommendation System's User Model

10. In the last 30 days, which parts of your identity were **not** well-represented in what TikTok thinks you like? Give three examples.

11. In the last 30 days, which parts of your identity were **not** well-represented in what TikTok thinks you like? Give three examples.

12. In the last 30 days, which parts of your identity were **not** well-represented in what TikTok thinks you like? Give three examples.

13. Look at your answers for content that TikTok frequently puts into your feed. You wrote "THEIR RESPONSE TO QUESTION (4) HERE". Give your best explanation of why TikTok puts this content into your feed. Write at least 2 sentences.

14. Look at your answers for content that TikTok frequently puts into your feed. You wrote "THEIR RESPONSE TO QUESTION (5) HERE". Give your best explanation of why TikTok puts this content into your feed. Write at least 2 sentences.

15. Look at your answers for content that TikTok frequently puts into your feed. You wrote "THEIR RESPONSE TO QUESTION (6) HERE". Give your best explanation of why TikTok puts this content into your feed. Write at least 2 sentences.

16. You wrote "THEIR RESPONSE TO QUESTION (10) HERE" as a part of your identity that is not well-represented. Please give your best explanation for why TikTok does not show you this content.

17. You wrote "THEIR RESPONSE QUESTION (11) HERE" as a part of your identity that is not well-represented. Please give your best explanation for why TikTok does not show you this content.

18. You wrote "THEIR RESPONSE QUESTION (12) HERE" as a part of your identity that is not well-represented. Please give your best explanation for why TikTok does not show you this content.

4.D.4 Private Mode

19. Suppose there was a "private mode" of TikTok, where you were not logged in and/or TikTok did not record which content you consume. What would you do differently on the platform, and what would you do the same way as currently? Please explain in 3-4 sentences.

4.D.5 User Content Curation

20. Which actions do you take to curate the content you are seeing in your TikTok feed? Please describe in 1-2 sentences.

21. Assume that you want to see **more sports content** in your feed in the future. Which of the following videos would you be **most likely** to engage with to reach this goal?

    (a) A beard-cutting tutorial

    (b) A song performance in a historic venue

    (c) A makeup tutorial

22. Assume that you want to see **more sports content** in your feed in the future. Which of the following videos would you be **most likely** to engage with to reach this goal?

    (a) A video with gossip about Kim Kardashian

(b) A weight loss motivational video

(c) A dog training video

23. Assume that you want to see **less makeup content** in your feed in the future. Which of the following videos would you be **most likely** to engage with to reach this goal?

    (a) A dog training video

    (b) An NFL breakdown, play-by-play

    (c) A romantic comedy clip

24. Assume that you want to see **less makeup content** in your feed in the future. Which of the following videos would you be **most likely** to engage with to reach this goal?

    (a) A feminist hip-hop clip

    (b) A weightlifting competition

    (c) A dance video

25. Describe in 3 sentences why you chose to answer the previous four questions as you did.

4.D.6 Demographics

27. What is your ethnicity?

    (a) White

    (b) Black of African American

    (c) American Indian or Alaska Native

    (d) Asian

    (e) Native Hawaiian or Pacific Islander

    (f) Other

28. What is the highest education level that you have achieved?

    (a) No schooling completed

    (b) Nursery school to 8th grade

    (c) Some high school, no diploma

    (d) High school graduate, diploma or the equivalent (for example: GED)

    (e) Some college credit, no degree

    (f) Trade/technical/vocational training

    (g) Associate degree

    (h) Bachelor's degree

(i) Master’s degree
(j) Professional degree
(k) Doctorate degree
(l) 2-year college graduate
(m) 4-year college graduate
(n) graduate degree
(o) post-graduate degree

29. What is your professional status?

    (a) Employed for wages
    (b) Self-employed
    (c) Out of work and looking for work
    (d) Out of work but not currently looking for work
    (e) A homemaker
    (f) A student
    (g) Military
    (h) Retired
    (i) Unable to work

30. What is your age?

    (a) 20 or younger
    (b) 20-29 years old
    (c) 30-39 years old
    (d) 40-59 years old
    (e) 60 years or older

31. Describe your gender identity

    (a) man
    (b) woman
    (c) Self-describe
    (d) Prefer not to say

Chapter 5

Regret Signals in Personalization

The contributions in this thesis up to now considered algorithmic demand, the likelihood that algorithms will choose different personalized experiences. The last two chapters in this thesis are considering two novel concerns about the interaction of algorithms and humans.

We start with an argument for handling behavioral inconsistency, in which case the stated and revealed preferences of the same person are inconsistent at different points in time.

5.1 Introduction

Actions that users take online might not match their intentions and may lead to undesired allocation of personalized experiences. Consider, for example, the behavior of doomscrolling, which the Merriam-Webster Dictionary defines as "*to spend excessive time online scrolling through news or other content that makes one feel sad, anxious, angry, etc.*" Why do users do what makes them feel sad, anxious, or angry? Why do users consume content if they later feel bad having done so? And what should personalized experiences look like that take into account the mismatch between intention and actions?

One explanation for doomscrolling is uncertainty: Users may not be aware that the content they choose makes them sad before they consume it. Another explanation is that users have self-control issues with content that looks engaging. In either case, the user will be able *after consuming the content* to state that they *regret* having consumed it. We present a data collection tool for retrospective signals of regret. This can be used to understand channels leading to regrettable behavior and to improve personalization. We close how regret signals can be used as a new frontier for consumer protection with algorithms.

Related Work Platforms serving personalized experiences already use retrospective evaluations for personalization. Goodrow (2021) describes how youtube.com uses and estimates *valued watch time*, which relies on retrospective evaluation of users whether their consumption of content was "valuable". Similarly, facebook.com uses surveys to personalize user feeds (Sethuraman, Vallmitjana, and Levin 2019; A. Gupta 2021), prompting users whether the content consumption was "worth their time". Our measurement tool differs in three dimensions, each important in its own right: (a) it leads to evaluation of content *outside* of the engineered environment of a website or app serving personalized experiences; (b) it explicitly elicits negative emotions that users have around their consumption; and (c) it is portable across platforms and can be performed based purely on data exports that are mandated under the European Union's General Data Protection Regulation and public Application Programming Interfaces.

Behavioral economics studies economic behavior outside of classically expected utility maximization and connects applied ethics with economic questions. We relate to several branches of literature in behavioral economics. A first is the debate on inner rational agents and the possibility of inferring *true* utility from behavior, see (Sugden 2004; Sugden 2018; Bernheim and Rangel 2009; Bernheim 2016; Sugden 2021). Depending on interpretation of regret data, regret data can be used to elicit inner rational preferences. More recently, computer scientists and economists have proposed that the inference of internal user states is a computational problem. Kleinberg et al. (2024) proposes to extract users' mental states from behavior and to take actions for users based on these signals, which would also benefit from regret data.

Other authors have considered alternative proxies for user utility. Agan et al. (2023) considers the amount of automaticity in two different features of facebook.com. Milli, Belli, and Hardt (2021) used several explicit feedback signals on twitter.com in addition to consumption signals to improve personalization. We add to this literature with a new data modality.

5.2 The Regret Data Collection Tool

We present a portable process to collect regret data for the example youtube.com. youtube.com is a video hosting service that serves personalized experiences on their main page, their apps, and as "watch next" recommendations. We focus on videos and "shorts" (videos of one minute or less, in a portrait format) on youtube.com, and reconstruct relevant information to help users recollect their previous consumption experience and express their regret, if any.

Figure 5.1 shows the interaction sequence. A user first retrieves usage data from takeout.google.com.[1] Next, they provide this data and their time zone to the tool. After processing the watch history, the tool shows users information on videos the users have watched and allows users to express which of the videos in their history they regret having watched.

To process the session information, we first split videos users watched into "sessions". We define a session as a sequence of videos that is an inclusion-maximal sub-sequence with the property that all videos in the sub-sequence start within 15 minutes of adjacent videos' start time. We present the user with a random subset of sessions with a minimum length of three videos and show users the first eight videos of a session or the whole session, whichever is shorter.

[1] The possibility of such data portability is mandated, at least in the European Union, under the General Data Protection Regulation's Art. 20.

Before starting to rate video-by-video, a session screen provides an overview of the session on which data will be elicited next. It contains the number of videos in the session, the start and end times of the session, and video thumbnails of all the videos watched. The user proceeds to state their regrets, video by video. We show a stylized depiction of the interface in A. Haupt and Curmei (2024) in Figure 5.2. Being given information on the video, such as the date and time when they started watching the video, its thumbnail, length, channel, and video, upload time, views, and video description, a user states whether they *regret* having watched the video, they *liked watching it*, they *do not remember* having watched the video, or *do not want to disclose*. They can input either through clicks or by using arrow buttons on their keyboard.

After inputting all their data, the user reviews their information before submitting it. The tool includes mechanisms for attention checks, is containerized, and stores regrets in a database.

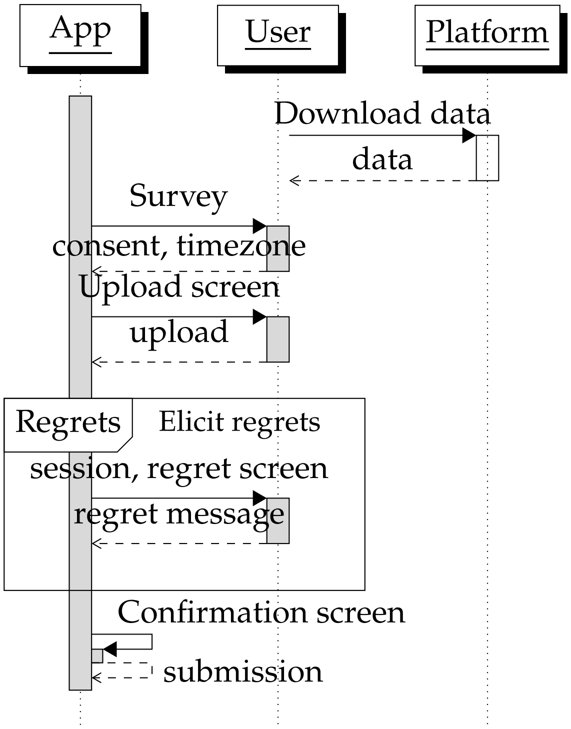


Figure 5.1: Regret data collection tool; sequence diagram

5.3 Uses of Regret Data

We close this chapter with examples of the use of regret data.

5.3.1 Regret Data for Providers of Personalized Experiences

Regret data can straightforwardly be incorporated into personalization algorithms. For example, instead of choosing content that approximately maximizes consumption prob-

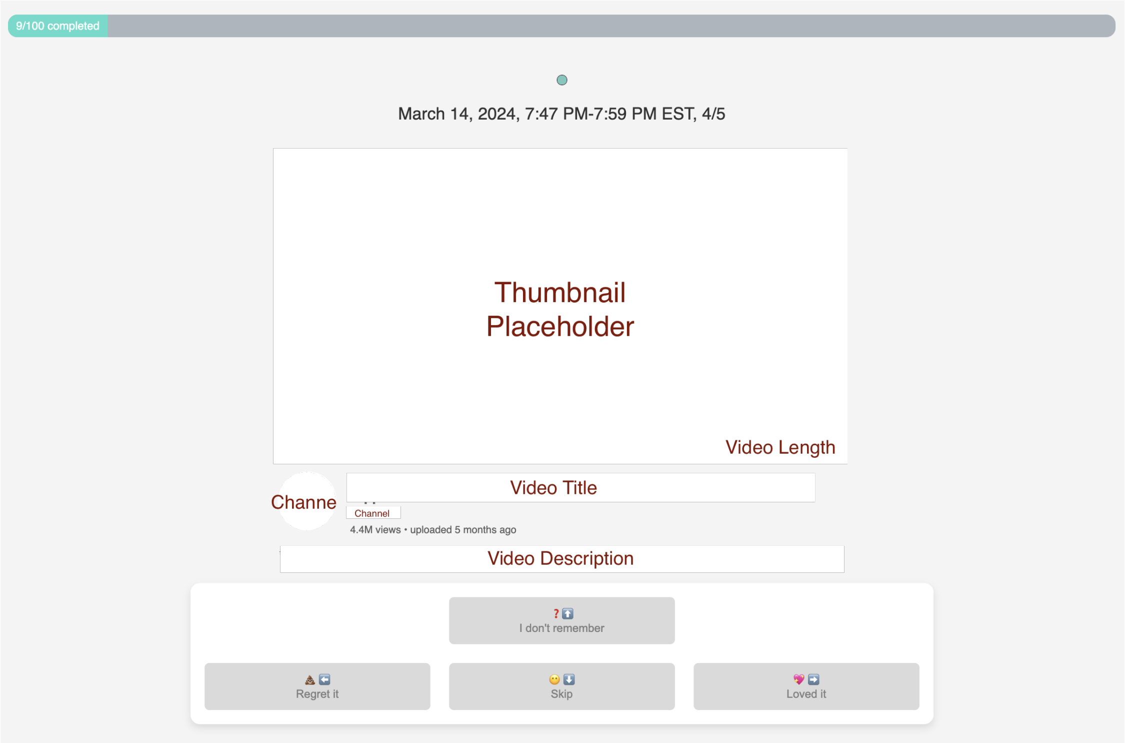


Figure 5.2: Regret data collection tool; video page

ability, content that is most likely to be consumed *and not regretted after the fact* can be selected.

A second use case is for web extensions filtering content and middleware (Fukuyama et al. 2020). For example, extensions could grey out content that is predicted to be regretted if consumed, or middleware could exclude content that is predicted to be regretted in what it serves to a user.

5.3.2 Regret Data for Regulators

As a second use, regret data can be used for consumer protection cases. In a future consumer protection regime, if a sample of users shows regret for a particular personalized experience, an investigation into whether the choice environment induces regrettable actions through deception or behavioral manipulation could be triggered.

5.3.3 Regret Data for Decision Support

Regret data can also be used by users themselves. Is regret correlated with features of a personalized experience, *e.g.*, does the user more frequently regret short videos or long videos, or is regret correlated with the sentiment of the video?

5.3.4 Regret Data for Social Scientific Research

Finally, regret data can advance behavioral science and applied microeconomic research.

When combined with other data, it allows researchers to better understand self-control issues versus learning dynamics and affective dimensions of consumption to disentangle different sources of inconsistent behavior.

Chapter 6

Contextual Privacy

The final chapter of this thesis considers the fundamental question of eliciting any preferences from a user: When is data elicitation justified? The chapter proposes *contextual privacy* as a formalization of Nissenbaum (2004)'s contextual integrity as a criterion of justified data elicitation. Only data that is directly relevant to a decision shall be elicited from an agent. We show that the information of multiple agents jointly, but not individually mattering for the computation of an outcome, renders contextually private elicitation impossible. This is the case for environments that deal with scarce resources, such as auctions or matching. When full privacy is impossible to achieve, we can achieve privacy for as many agents as possible, allocating the privacy burden to some to protect the privacy of others.

6.1 Introduction

In standard mechanism design, a designer elicits reports of agents' private information in order to determine the outcome of a social choice rule. Ex-post, in an incentive compatible direct revelation mechanism, the designer learns *all* of agents' private information—typically more than is necessary for computing the rule. For example, in a sealed-bid second price auction, the designer learns all losing bids exactly, even though it is only necessary to know that all losing bids fall below the second highest. The designer also learns the winner's bid exactly, even though it is only necessary to know that the winning bid is above the second highest.

At the same time, there are several reasons why it may be desirable that the designer not learn "too much" about agents. First, gaining excess knowledge of agents' private information could expose the designer to legal liabilities or political risk. For example, as recounted in McMillan (1994), a second-price auction for spectrum licenses in New Zealand had a "political defect": "by revealing the high bidder's willingness to pay, the auction exposed the government to criticism" (McMillan 1994). The government would have benefited from a design that ensured it only learned what was strictly needed—with the idea that, as the adage goes, what they don't know can't hurt them. Second, if agents have privacy concerns, they may be reluctant to reveal their private information even if the appropriate allocative incentives are in place. Agents may worry that their private information could be used against them in subsequent interactions with the designer or third parties (Rothkopf, Teisberg, and Kahn 1990; Ausubel 2004). Indeed, a 2022 lawsuit alleged that Google stored losing bids in their second-price auctions for advertising, and used this information against advertisers in future interactions.[1]

In this paper, we study how mechanism designers can limit the superfluous information they learn. In our set-up, when a designer commits to a social choice rule, they also choose a dynamic protocol for eliciting agents' information (or "types"). These dynamic protocols allow the designer to learn agents' private information gradually, ruling out possible type profiles until they know enough to compute the outcome of the rule. The key idea we introduce is a *contextual privacy violation*. A protocol produces a contextual privacy violation for a particular agent if the designer learns a piece of their private information that is superfluous, *i.e.* it might not be necessary for computing the outcome. We study protocols that are on the frontier of contextual privacy and implementation—that is, we study a setwise order based on contextual privacy violations, and look for maximal elements in this order. For some choice rules, it will be possible to find protocols that produce no violations for any agent at any type profile—we call these protocols *contextually private*.

We define protocols in a way that allows us to accommodate a broad range of environments. A protocol is composed of queries. A query can be directed to one agent—for example, a designer may ask one agent "Is your type above x?" Or, a query can be directed to multiple agents—*e.g.*, the designer could ask "How many agents have a type above x?" A query can also be more complex than a simple threshold, for example, the designer could ask "What is the average type among agents with types between x and x'?" The details of the environment may dictate the format of queries the designer is able to ask—we call the set of available queries the *elicitation technology*.

An elicitation technology represents *how* the designer can learn about agents' messages, and by inverting their strategies, their private information. One possible elicitation technology is a "trusted third party." If there is a trusted third party, the designer can delegate information retrieval to this third party, and ask the third party to report back only what is needed to compute the outcome. So, under this trusted third party elicitation technology, all choice rules can trivially eliminate all contextual privacy violations. Cryptographic techniques like secure multi-party computation and zero-knowledge proofs are elicitation technologies that similarly trivialize contextual privacy.[2]

But there are many environments in which the social or technological environment will neither permit the use of a trusted third party nor of trusted cryptographic tools. Advanced cryptography may be excessively costly in terms of time, money or computational power.[3] In addition, sophisticated cryptographic mechanisms require sophisticated participants—

[1] In particularly, the lawsuit *State of Texas v. Google* alleged that "Google induced advertisers to bid their true value, only to override pre-set AdX floors and use advertisers' true value bids against them... generat[ing] unique and custom per-buyer floors depending on what a buyer had bid in the past."

[2] For a survey of cryptographic protocols for sealed-bid auctions, see Alvarez and Nojoumian (2020).

[3] Even if possible, some sophisticated solutions may be wasteful—in one of the earliest large-scale uses of secure multi-party computation, a double auction with sugar beet farmers in Denmark, designers wondered "if the full power of multiparty computation was actually needed," or if a simpler implementation guided by a weaker privacy criterion may have sufficed (Bogetoft et al. 2009).

if participants do not understand how their information is kept private, their privacy concerns may not be alleviated.[4]

We focus on protocols founded on minimal assumptions regarding trust and comprehension. Specifically, for the most design-relevant portions of the paper, we restrict attention to *individual elicitation* technologies. In an individual elicitation protocol, the designer is limited to queries directed at individual agents, sequentially learning about this agent's type. Individual elicitation protocols neatly expose the privacy properties of a given protocol: for an agent to understand what the designer has learned about her, she merely has to recall her responses to the designer's questions. In other words, such protocols are unmediated.[5] They have the convenient property that the designer knows a piece of an agents' private information if and only if the agent said it.

We now present a brief overview of the paper. In Section 6.2, we articulate our formal framework and present the key definitions of the paper. We first must formally define the dynamic messaging game played by agents, and the elicitation technologies that limit what the designer learns about agents' messages. Then, we introduce the central definitions: contextual privacy violations, the contextual privacy order (an inclusion order based on violations), and maximally contextually private protocols (maximal elements in the inclusion order).

In Section 6.3, we study how specific properties of choice rules tend to lead to contextual privacy violations. In our first results, Proposition 6.1 and Proposition 6.2, we provide characterizations of choice rules that fully avoid contextual privacy violations under an arbitrary fixed elicitation technology, and under individual elicitation technologies, respectively. These abstract characterizations lead us to a more intuitive insight, Theorem 6.1, which says that under the restriction to individual elicitation protocols, any time there is some group of agents who are *collectively pivotal* but there is no agent who is *individually pivotal*, there must be a contextual privacy violation and the designer must choose whose privacy to protect.[6] Propositions 6.3-6.7 show how this conflict between collective and individual pivotality arises in common choice rules in environments with and without transfers.

In Section 6.4, we take the perspective of a privacy-conscious designer who wants to implement a choice rule through a maximally contextually private protocol. Our discussion yields two central insights. First, maximally contextually private protocols involve a deliberate decision about the *protection set*, *i.e.* a set of agents whose privacy

[4] A recent survey shows that only 61 % of WhatsApp's users believe the company's claim that their messages are end-to-end encrypted (Alawadhi 2021).

[5] This minimal assumption stands in contrast to more substantive assumptions we could make. For example, if the designer's elicitation technology allows them to count the number of agents whose type satisfies a certain property without learning *whose* type satisfies that property, the agents must trust the designer's use of anonymization techniques. If the designer's elicitation technology enables the secret sharing behind secure multiparty computation, agents must grasp the idiosyncratic guarantees and computational assumptions of the particular secure multiparty computation protocol in use.

[6] One can derive as a special case of Theorem 6.1 a result from the cryptography literature known as the Corners Lemma (Chor and E. Kushilevitz 1989; Chor, Geréb-Graus, and Eyal Kushilevitz 1994). This result has been used to show that the second-price auction does not permit a decentralized computation protocol (Brandt and Sandholm 2005) that satisfies *unconditional privacy*, compare Chor and E. Kushilevitz (1989) and Milgrom and Segal (2020).

ought to be protected if possible. Second, given that protection set, maximally contextually private protocols delay asking questions to protected agents as much as possible.

To get to these insights, we first present a key theoretical result of the paper that helps us reason about the contextual privacy-implementation frontier. Theorem 6.2 shows that for choice rules defined on ordered type spaces, it is without loss for such a designer to consider only *bimonotonic* protocols. Formally, the result is a representation theorem, showing that every protocol is contextual privacy equivalent to a bimonotonic protocol. A bimonotonic protocol consists of *threshold queries* which, for each agent, are monotonically increasing or decreasing in the threshold.

We then turn to the special case of the second-price auction rule as an instructive and practically relevant example—we see how maximally contextually private protocols for the second-price auction rule choose a set of agents to protect, and delay asking questions to the protected agents. In Theorem 6.3, we use the representation theorem (Theorem 6.2) to derive two maximally contextually private protocols: the *ascending-join* and the *overdescending-join* protocols. In the ascending-join protocol, which is a variant of the familiar ascending or "English" auction protocol, the designer begins by conducting an ascending protocol with just two agents. Whenever one agent drops out, another agent "joins" the protocol at the going threshold. The overdescending-join protocol is analogous, and related to the overdescending protocol introduced in Harstad (2018). The ascending-join protocol protects the winner, the overdescending-join protocol protects the losers, and both protocols avoid violations by delaying questions to agents in the protected set as much as possible.

Under the restriction to individual elicitation, protocols induce a well-defined extensive-form game. In Section 6.5, we "check" the incentive properties of the maximally contextually private protocols for the second-price auction rule described in Section 6.4. The ascending-join and overdescending-join protocols have implementations in dominant strategies, with the former satisfying the stronger requirement of obvious strategyproofness. This section relies on prior results regarding the strategic properties of personal clock-auctions (Li 2017).

Finally, we discuss related literature in Section 6.6 and conclude in Section 6.7.

Proofs omitted from the main text are in Section 6.A.

6.2 Model

There is a set $N = \{1, 2, \dots, n\}$ of agents with private information (or "types") θ_i distributed according to $F_i \in \Delta(\Theta)$, where Θ is a finite type space. We denote by $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n) \in \Theta^n = \boldsymbol{\Theta}$ a profile of agents' types.[7] Agents have utility functions u over outcomes in X which may depend on their private types, $u_i : \Theta \times X \to \mathbb{R}$. All primitives of the model besides the realized type profile $\boldsymbol{\theta}$ are common knowledge.

A designer implements a deterministic social choice function $\phi \colon \boldsymbol{\Theta} \to X$ through a dynamic protocol in which agents repeatedly send messages m from a message space M.

[7] We follow the usual convention of referring to partial type profiles excluding an agent i by $\boldsymbol{\theta}_{-i} = (\theta_j)_{j \in N \setminus \{i\}}$. We also sometimes extend this notation to refer to partial type profiles excluding sets of agents $A \subseteq N$, *i.e.* $\boldsymbol{\theta}_{-A} = (\theta_j)_{j \in N \setminus A}$.

6.2.1 Protocols.

The protocol consists of rounds. In each round, all agents submit a message $m \in M$ from a finite message space, with message profile $\boldsymbol{b} \in \mathbf{M} := M^n$. The designer chooses a partition $\mathcal{S} \subset 2^{\mathbf{M}}$ and observes which partition cell $S \subset \mathcal{S}$ the message profile lies in.

Not all partitions are admissible. The designer must choose partitions from an *elicitation technology*, a set $\mathfrak{S} \subseteq 2^{2^{\mathbf{M}}}$ of admissible partitions. Most of this paper considers a particular elicitation technology that mimics communication with one agent at a time. The *individual elicitation technology,* denoted $\mathfrak{S}_{\text{IE}}$, contains n partitions:

$$\mathcal{S}_i = \{\{\boldsymbol{b} \in \mathbf{M} : m_i = m\} : m \in M\}. \tag{6.1}$$

The partition cells of $\mathcal{S}_i$ differ only in one agent i's message. So, we can interpret protocols that exclusively use the individual elicitation technology as protocols in which the designer queries one agent at a time. These protocols are unmediated in the sense that the designer need not *commit to forget* anything about agent i's submitted report.

To represent salient features of the designer's social and technological environment one could consider a variety of elicitation technologies. In Section 6.D, we consider the *count elicitation technology,* under which a designer can observe the number of agents whose messages satisfy a given property, without learning the identities of agents who send a message with the given property. This elicitation technology corresponds to the common practice of anonymized elicitation (for example, many political elections use a secret ballot and online auctions permit anonymous bidding). Or, more formally, these technologies represent the designer's commitment to forget the labels of submitted messages.

Each protocol—a series of queries that may adaptively depend on prior queries—can be represented as a directed rooted tree. The tree's nodes correspond to histories of chosen partitions and observed partition cells. We denote the root node by r and terminal nodes by $z \in Z$. The set of all nodes of the tree is V and the set of edges is E. We will also write $P = (V, E)$ for a protocol. We encode the choice of partition at a node v through a function $s_v : \mathbf{M} \to \text{children}(v) \subset V$. Each child of v corresponds to one of the partition cells. It is then equivalent to using an elicitation technology $\mathfrak{S}$ such that,

$$\{s_v^{-1}(w) \subseteq \mathbf{M} : w \in \text{children}(v)\} \in \mathfrak{S}.$$

In this formalism, nodes v encode the full history of the chosen partitions and observed cells. We refer to a protocol with elicitation technology $\mathfrak{S}$ as an $\mathfrak{S}$-protocol, and denote by $\mathcal{P}_{\mathfrak{S}}$ the set of all protocols with elicitation technology $\mathfrak{S}$. Table 6.1 compiles notation.

6.2.2 Strategies.

At non-terminal nodes, agents submit messages according to deterministic strategies $\sigma_i : (V \setminus Z) \times \Theta \to M$, with strategy profiles denoted $\boldsymbol{\sigma} := (\sigma_1, \sigma_2, \ldots, \sigma_n)$. That is, depending on their type θ_i, an agent chooses a message to send at each non-terminal node $v \in V$. We will write $\boldsymbol{\sigma}_v : \boldsymbol{\Theta} \to \mathbf{M}$ to refer to the strategy profile evaluated at node v.

Nodes $v \in V$ may be identified with the *information available to the designer*, which is the set of all type profiles that are consistent with the observed partition cells so far,

$$\mathbf{\Theta}_v := \bigcap_{(u,w)\in \mathrm{path}(v)} \boldsymbol{\sigma}_u^{-1}(s_u^{-1}(w)),$$

where, path(v) is the set of edges from the root node r to the node v. Protocols can be equivalently represented as $((\mathbf{\Theta}_v)_{v\in V\setminus\{r\}}, \phi)$ and $((s_v)_{v\in V\setminus Z}, \phi)$.[8]

We inductively define the outcome $P(\boldsymbol{\sigma}(\boldsymbol{\theta}))$ of protocol P under type profile $\boldsymbol{\theta}$ and strategies $\boldsymbol{\sigma}$. The induction starts with the root node r. For any non-terminal node $v \in V \setminus Z$, the successor node is

$$s_v(\sigma_1(\theta_1, v), \sigma_2(\theta_2, v), \dots, \sigma_n(\theta_n, v)) \in V.$$

At a terminal node $z \in Z$, we choose the outcome x associated with z. We say that P is a protocol *for* choice rule ϕ with strategies $\boldsymbol{\sigma}$ if

$$P(\boldsymbol{\sigma}(\boldsymbol{\theta})) = \phi(\boldsymbol{\theta}).$$

If there is an $\mathfrak{S}$-protocol for a choice rule ϕ with some strategies $\boldsymbol{\sigma}$, we say that the choice rule is $\mathfrak{S}$-*computable*. We will also sometimes call $(P, \boldsymbol{\sigma})$ a protocol.

Name	Sets	Representative Element
Agents	$\{1, \dots, n\}$	i
Type profiles	$\mathbf{\Theta} = \Theta^n$	$\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$
Message profiles	$\mathbf{M} = M^n$	$\boldsymbol{b} = (m_1, \dots, m_n)$
Queries	$\mathfrak{S}$	$\mathcal{S}$, s_v
Message partition cells	$\mathcal{S}$	S
Protocols	$\mathcal{P}_{\mathfrak{S}}$	$P = (V, E)$
Nodes	V	v, w
Edges	E	$e = (v, w)$
Terminal nodes of protocol P	Z	z
Information at node v	$\mathbf{\Theta}_v$	
	Relations	
Agent i prefers outcome x to x'		$x \succ_i x'$
Protocol P is more contextually private than P' for ϕ		$P \succ_\phi P'$
Type θ is succeeded by θ' in the type space		$\theta' = \mathrm{succ}(\theta)$ or $\theta = \mathrm{pred}(\theta')$

Table 6.1: Notation for Protocols and Relations

6.2.3 Contextual Privacy Violations.

Type profiles $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{\Theta}$ lead to different terminal nodes if and only if there is a non-terminal node that *distinguishes* them, *i.e.* there are sibling nodes $v, v' \in V$ such that $\boldsymbol{\theta} \in \mathbf{\Theta}_v$ and $\boldsymbol{\theta}' \in \mathbf{\Theta}_{v'}$. We then also say that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are distinguished *at* $v \in V$.

[8] That is, the set of protocols $\mathcal{P}_{\mathfrak{S}}$ represented as trees with node labels s_v for non-terminal nodes is isomorphic to the set $\mathcal{P}_{\mathfrak{S}}$ with node labels $\mathbf{\Theta}_v$.

The main definition of this article is the *contextual privacy violation*. A contextual privacy violation is a piece of agent i's private information learned by the designer that did not play a role in determining the outcome at a given type profile $\boldsymbol{\theta}$. All of the key concepts of the paper employ this definition.

Definition 6.1 (Contextual Privacy Violation)**.** Let P be a protocol for ϕ with strategies $\boldsymbol{\sigma}$. We say that P and $\boldsymbol{\sigma}$ produce a ϕ*-contextual privacy violation* for agent i at $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i}) \in \boldsymbol{\Theta}$ if there is a type $\theta_i' \in \Theta$ such that

$$(\theta_i, \boldsymbol{\theta}_{-i}) \text{ and } (\theta_i', \boldsymbol{\theta}_{-i}) \text{ are distinguished, yet } \phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i}).$$

Otherwise, we say agent i's contextual privacy is *preserved* at $\boldsymbol{\theta}$. We denote by $\Gamma(P, \boldsymbol{\sigma}, \phi) \subseteq N \times \boldsymbol{\Theta}$ the set of contextual privacy violations produced by $P, \boldsymbol{\sigma}$ and ϕ. If the strategy is clear from the context, we will also write $\Gamma(P, \phi)$ for brevity.

In other words, there is a contextual privacy violation for agent i under protocol P and choice rule ϕ at type profile $\boldsymbol{\theta}$ if the designer can tell apart types θ_i, θ_i' but θ_i and θ_i' lead to the same outcome under ϕ, holding fixed $\boldsymbol{\theta}_{-i}$.

An agent's privacy is preserved at a realized type profile $\boldsymbol{\theta}$ if the private information she reveals is necessary for computing the outcome. This notion of preservation complies with the principle of *purpose limitation* laid out in the European Union's General Data Protection Regulation (GDPR), implemented in 2018.[9] More broadly, contextual privacy preservation aligns with the influential theory of privacy as *contextual integrity* (Nissenbaum 2004), which states that an information flow (*e.g.*, from the agent to the designer) is private if it respects contextual expectations and norms (*e.g.*, that the designer learns agent's private information *only* for the purposes of computing the choice rule).

Contextual privacy can also be used to formulate other meaningful desiderata—for example, the idea that queries should have only *legitimate explanations*.[10] Imagine that the designer is asked to explain why they must learn something about agent i, *i.e.* they are asked why they need to know that agent i's type is θ_i and not θ_i'. A *legitimate explanation* $\boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}^{n-1}$ is a profile of other agents' types that explains why they need this information. That is, an *explanation* is a profile of other agents' types $\boldsymbol{\theta}_{-i}$ under which the designer claims they will need to distinguish between types θ_i and θ_i'—an explanation $\boldsymbol{\theta}_{-i}$ is *legitimate* if the outcome of the choice rule would be different depending on the information distinguished about agent i. There is no contextual privacy violation for agent i, that is $\Gamma(P, \phi) \cap (\{i\} \times \boldsymbol{\Theta}) = \emptyset$, when all possible explanations are legitimate. Or, put in the contrapositive, there is a contextual privacy violation when there is some explanation that is not legitimate.

In Section 6.E, we explore two modifications of contextual privacy violations that may track different design goals: individual contextual privacy and group contextual

[9] The principle of *purpose limitation* appears in Article 5(1) of the GDPR, and reads: "Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes."

[10] The idea of a legitimate explanation parallels Akbarpour and Li (2020)'s notion of an *innocent explanation* which is used to define credibility: in their setting, an innocent explanation is a profile of other agents' types $\boldsymbol{\theta}_{-i}$ that can explain the observed outcome.

privacy. These variations highlight connections to other desiderata such as non-bossiness (Satterthwaite and Sonnenschein 1981; Pycia and Raghavan 2021) and (strong) obvious strategyproofness (Li 2017; Pycia and Troyan 2023).

The goal of the privacy-conscious market designer is to choose a protocol that is, in some sense, "better," from a privacy perspective, than other protocols that achieve the same outcome. In order to study this goal, we define the *contextual privacy order*.

Definition 6.2 (Contextual Privacy Order)**.** A protocol P for ϕ with strategies $\boldsymbol{\sigma}$ is *more contextually private* than another protocol P' for ϕ under strategies $\boldsymbol{\sigma}$, denoted $(P,\boldsymbol{\sigma}) \preceq_{\phi} (P',\boldsymbol{\sigma}')$ if $\Gamma(P,\boldsymbol{\sigma},\phi) \subseteq \Gamma(P',\boldsymbol{\sigma}',\phi)$. A protocol $(P,\boldsymbol{\sigma})$ for ϕ, $P \in \mathcal{P}_{\mathfrak{S}}$, is $\mathfrak{S}$*-maximally contextually private* if there is no $P' \in \mathcal{P}_{\mathfrak{S}}$ and no strategy $\boldsymbol{\sigma}'$ such that $\Gamma(P',\boldsymbol{\sigma}',\phi) \subsetneq \Gamma(P,\boldsymbol{\sigma},\phi)$.

The contextual privacy order is an inclusion order on the set of contextual privacy violations that allows us to compare two protocols P and P' for a given choice rule ϕ.[11] We call $\preceq_{\phi}$-maximal objects within $\mathcal{P}_{\mathfrak{S}}$ also *maximally contextually private* if the elicitation technology, strategies, and social choice function are clear from the context.

The contextual privacy order captures another principle of the EU's GDPR law, known as *data minimisation.*[12] This order takes the pragmatic approach of locating protocols that violate contextual privacy as little as possible while still computing the choice rule. Maximally contextually private protocols are those that trade off contextual privacy and computability in a Pareto optimal way, and thus can be seen as lying on the *privacy-implementation frontier*.

The contextual privacy order is robust in the following sense: if a protocol P is more contextually private than a protocol P', it elicits superfluous information from fewer agents at *any* type profile $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. That is, it is more private for any prior over the type profile F.

For the extreme case in which a maximally contextually private protocol leads to no contextual privacy violation at all, *i.e.* there exists a P such that $\Gamma(P,\phi)$, we also call the choice rule ϕ *contextually private*.

6.2.4 Incentives.

As the agents' strategies $\sigma_i : \Theta \times V \to M$ condition on the full state of the protocol at every node, dynamic incentives may be complex under arbitrary elicitation technologies. As our main goal in this article is to introduce and study the demands of contextual privacy, we for the most part isolate contextual privacy concerns from incentive compatibility concerns. We give necessary conditions for contextual privacy that consider any (incentive compatible or not) strategies $\boldsymbol{\sigma}$ and we verify dynamic incentive compatibility for the

[11] Segal (2007) and Mackenzie and Zhou (2022) consider a related order, the *relative informativeness*, on the information revealed by different extensive-form implementations of choice rules. In Mackenzie and Zhou (2022), the relative informativeness order is invoked to illustrate that menu mechanisms can be less informative than direct revelation mechanisms, while in Segal (2007), the order is employed to identify the communication costs of choice rules.

[12] Article 5(1) of the EU's GDPR law lays out the principle of data minimization as follows: "Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".

protocols we construct in Section 6.5 and Section 6.B. We defer our definitions of dynamic incentive compatibility to Section 6.5.

6.3 How Contextual Privacy Violations Arise

We first study how contextual privacy violations depend on the social choice rule a designer wishes to implement. We do this in two ways. First, we provide a general characterization of choice rules that admit a contextually private protocol—*i.e.* choice rules for which it is possible to find a protocol that produces *no* contextual privacy violations at any type profile. We give such characterizations in Section 6.3.1—one characterization is fully general, for an arbitrary fixed class of elicitation technologies, and the other characterization is for the restriction to individual elicitation technologies. Second, we use these characterizations to derive a simple illustration of how and where contextual privacy violations tend to arise under individual elicitation in common choice rules. That is, in Section 6.3.2, we see that whenever there is *collective pivotality but no individual pivotality*, there will be a violation. This principle helps us to understand not just whether contextual privacy will be violated but *where* contextual privacy fails in common choice rules.

6.3.1 Two Characterizations of Contextually Private Choice Rules.

We first characterize $\mathfrak{S}$-contextually private choice rules, for an arbitrary fixed class of elicitation technologies $\mathfrak{S}$ in Proposition 6.1. Then we offer in Proposition 6.2 a characterization that is specialized to the individual elicitation technology $\mathfrak{S}_{\mathrm{IE}}$.

Proposition 6.1. *An $\mathfrak{S}$-computable choice rule ϕ is $\mathfrak{S}$-contextually private if and only if there does not exist a subset of type profiles $\hat{\boldsymbol{\Theta}} \subseteq \boldsymbol{\Theta}$ such that*

(i) $\phi|_{\hat{\boldsymbol{\Theta}}}$ is non-constant, and

(ii) for every partition $\mathcal{S} \in \mathfrak{S}^*$ such that none of the partition cells $S \in \mathcal{S}$ contains $\hat{\boldsymbol{\Theta}}$, there are type profiles $(\theta_i, \boldsymbol{\theta}_{-i}), (\theta_i', \boldsymbol{\theta}_{-i}) \in \hat{\boldsymbol{\Theta}}$ that lie in different partition cells of $\mathcal{S}$.

Here, $\mathfrak{S}^* \subseteq 2^{2^{\boldsymbol{\Theta}}}$ is the set of revealed partitions. *Define $\mathcal{S}^* \in \mathfrak{S}^*$ if there is a partition $\mathcal{S} \in \mathfrak{S}$ and functions $f_1, f_2, \ldots, f_n \colon \Theta \to M$ such that every $S^* \in \mathcal{S}^*$ is the preimage under $(f_1, f_2, \ldots, f_n) \colon \boldsymbol{\Theta} \to \mathbf{M}$ of some $S \in \mathcal{S}$, $S^* = (f_1, f_2, \ldots, f_n)^{-1}(S)$.*

In addition, if there is such a $\hat{\boldsymbol{\Theta}}$ then any protocol for ϕ must produce a contextual privacy violation for some type profile in $\hat{\boldsymbol{\Theta}}$, that is, it must be that for any $P \in \mathcal{P}_{\mathfrak{S}}, \Gamma(P, \boldsymbol{\sigma}, \phi) \cap (\hat{\boldsymbol{\Theta}} \times N) \neq \emptyset$.

Note that this characterization is purely about the properties of a choice rule ϕ, and suggests that whether there is an admissible protocol P that does not produce a contextual privacy violation can be understood in terms of how the choice rule acts on subsets of the space of type profiles. This result is valuable because it tells us that for an arbitrary elicitation technology, small "counterexamples" can entirely characterize contextual privacy. For example, in the count elicitation technology discussed in the model section and

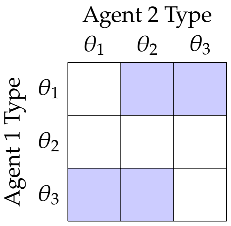


Figure 6.1: Illustration of Inseparable Types with $n = 2, \hat{\boldsymbol{\Theta}} = \{\theta_1, \theta_2, \theta_3\}^2$. Shaded regions represent outcome x under ϕ. For agent 1, $\theta_3 \sim_{1,\phi,\hat{\boldsymbol{\Theta}}} \theta_1$. For agent 2, $\theta_1 \sim_{2,\phi,\hat{\boldsymbol{\Theta}}} \theta_2 \sim_{2,\phi,\hat{\boldsymbol{\Theta}}} \theta_3$.

in detail in Section 6.D, a counterexample of four type profiles forces a contextual privacy violation for some agent.

That said, at this level of generality, this theorem gives little insight into *which* subsets of the space of type profiles $\hat{\boldsymbol{\Theta}}$ are likely to produce a violation. Checking whether conditions (i) and (ii) hold might require the consideration of all subsets of the space of type profiles $\hat{\boldsymbol{\Theta}}$ as well as any partition $\mathcal{S} \in \mathfrak{S}^*$. We now show that under a restriction to individual elicitation protocols, we can derive a more insightful characterization. This characterization relies on the notion of *type inseparability*.[13] Roughly, two types for an agent i are *inseparable* if the designer cannot distinguish between them without violating contextual privacy.

Definition 6.3 (Inseparable Types)**.** For a social choice function ϕ, call two types θ_i, θ'_i for an agent i *directly inseparable* on $\hat{\boldsymbol{\Theta}}$ under ϕ, denoted $\theta_i \sim'_{i,\phi,\hat{\boldsymbol{\Theta}}} \theta'_i$ if there exists $\boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}_{-i}$ such that $(\theta_i, \boldsymbol{\theta}_{-i}), (\theta'_i, \boldsymbol{\theta}_{-i}) \in \hat{\boldsymbol{\Theta}}$, and

$$\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta'_i, \boldsymbol{\theta}_{-i}).$$

Denote the transitive closure of $\sim'_{i,\phi,\hat{\boldsymbol{\Theta}}}$ by $\sim_{i,\phi,\hat{\boldsymbol{\Theta}}}$. If $\theta_i \sim_{i,\phi,\hat{\boldsymbol{\Theta}}} \theta'_i$, call θ_i and θ'_i *inseparable* for i. We denote equivalence classes under $\sim_{i,\phi,\hat{\boldsymbol{\Theta}}}$ by $[\theta]_{i,\phi,\hat{\boldsymbol{\Theta}}}$.

To build further intuition for this definition, Figure 6.1 gives an illustration of inseparable types. The 3×3 grid represents a subset of the type space in a setting where there are two agents ($n = 2$). The shaded regions of the grid represent type profiles for which the outcome under ϕ is a particular outcome $x \in X$. Regions of the grid that are not shaded lead to arbitrary (other than x) outcomes under ϕ. On $\hat{\boldsymbol{\Theta}}$, all of agent 2's types are inseparable. To see this, note that θ_1 and θ_2 are directly inseparable—they lead to the same outcome x when agent 1's type is fixed at θ_3. Furthermore, for agent 2, θ_2 and θ_3 are directly inseparable, since they lead to the same outcome x when agent 1's type is fixed at θ_1. So, since inseparability is transitive, θ_1, θ_2 and θ_3 are all inseparable for agent 2.

In short, when a choice rule ϕ requires separating inseparable types there is a contextual privacy violation under $\mathfrak{S}_{\text{IE}}$.

[13] Our concept of *inseparability* parallels the concept of *forbidden matrices* used in the work on 2-agent decentralized computation Chor and E. Kushilevitz (1989) and Chor, Geréb-Graus, and Eyal Kushilevitz (1994).

Proposition 6.2. *A choice function ϕ fails to be $\mathfrak{S}_{\text{IE}}$-contextually private if and only if there exists some cylinder set $\hat{\mathbf{\Theta}} = \bigtimes_{i=1}^{n} \hat{\Theta}_i$ such that (i) $\phi|_{\hat{\mathbf{\Theta}}}$ is non-constant and (ii) for all agents i and all $\theta_i, \theta'_i \in \hat{\mathbf{\Theta}}_i$, θ_i and θ'_i are inseparable with respect to $\hat{\mathbf{\Theta}}$ and ϕ.*

In addition, it must be that for any protocol $(P, \boldsymbol{\sigma}) \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$, one of the agents i that has at least two inseparable types must incur a privacy violation, that is, $\Gamma(P, \boldsymbol{\sigma}, \phi) \cap \{(i, \boldsymbol{\theta}) \in N \times \mathbf{\Theta} \mid \exists \theta'_i \neq \theta_i : (\theta_i, \boldsymbol{\theta}_{-i}), (\theta'_i, \boldsymbol{\theta}_{-i}) \in \hat{\mathbf{\Theta}}\} \neq \emptyset$.

The proofs of both characterizations Proposition 6.1 and Proposition 6.2 are similar. For necessity, we consider the first query that distinguishes type profiles from $\hat{\mathbf{\Theta}}$ which must exist due to the assumed non-constancy of ϕ restricted to $\hat{\mathbf{\Theta}}$. The conditions on $\hat{\mathbf{\Theta}}$ in (ii) require that this query must lead to a contextual privacy violation. The proof of sufficiency is constructive.

These characterizations allow us to analyze the contextual privacy of choice rules, albeit in a way that does not directly offer insight into why contextual privacy fails. In the next subsection, we will leverage Proposition 6.2 to derive a more intuitive result that shows how contextual privacy violations arise in many common choice rules.

6.3.2 Collective and Individual Pivotality.

The main result in this section allows us to identify the type profiles at which contextual privacy violations arise. This result shows that if there is a type profile where agents are collectively pivotal, but no agent is individually pivotal, then there will be a violation for some agent at that type profile.

Theorem 6.1. *Let ϕ be a social choice function, and consider a type profile $\boldsymbol{\theta} \in \mathbf{\Theta}$. If for any subset $A \subseteq N$ of agents, and types $\theta'_i \in \Theta$ for agents $i \in A$,*

$$\phi(\boldsymbol{\theta}_A, \boldsymbol{\theta}_{-A}) \neq \phi(\boldsymbol{\theta}'_A, \boldsymbol{\theta}_{-A}) \qquad \text{(collective pivotality)}$$

and for all $i \in A$

$$\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta'_i, \boldsymbol{\theta}_{-i}) \qquad \text{(no individual pivotality)}$$

then for any individual elicitation protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$, there exists an agent $i \in A$ whose contextual privacy is violated at $\boldsymbol{\theta}$, i.e. $A \times \{\boldsymbol{\theta}\} \cap \Gamma(P, \phi) \neq \emptyset$.

This theorem tells us that a social choice function will produce a contextual privacy violation under any $\mathfrak{S}_{\text{IE}}$-protocol at a type profile $\boldsymbol{\theta}$ if there is a set of agents that *collectively* are pivotal for the outcome, but none of them is individually pivotal. This is intuitive, as the individual elicitation technology needs to query one agent at a time. So, a situation of collective pivotality without individual pivotality suggests that the designer must ask some agent a question that may not matter for the outcome, but had to ask that question in case it did matter.

In the remainder of this section, we discuss instances in which collective pivotality vs. individual pivotality arise in common choice rules. We consider specific choice rules in three domains: assignment, auctions and voting.

Table 6.2: Contextual Privacy Violations of Choice Rules Under Individual Elicitation Technology

Choice Rule ϕ	$\inf_{P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}} \Gamma(P, \phi)$	
Assignment (Section 6.3.2)		
Serial Dictatorship Rule	$= \emptyset$	Section 6.B.1
Efficient and IR Rules (House Assignment)	$\neq \emptyset$	Proposition 6.3
Stable Rules (Matching with Priorities)	$\neq \emptyset$	Proposition 6.4
Auctions (Section 6.3.2)		
First-price Auction Rule	$= \emptyset$	Section 6.B.2
Second-price Auction Rule	$\neq \emptyset$	Proposition 6.5
Efficient Double Auction	$\neq \emptyset$	Proposition 6.6
Voting (Section 6.3.2)		
Generalized Median Voting Rule	$\neq \emptyset$	Proposition 6.7

Application to Assignment.

In the assignment domain, we fix a set C of objects. The set of outcomes is $X = 2^{N \times C}$.

In the standard object assignment setting, agents may receive at most one object, and agents have ordinal preferences over objects, which are private information. So agents' types $\boldsymbol{\theta} \in \Theta$ are preference orders on C where $\succeq_i$ refers to agent i's preference ordering.

In the classic *serial dictatorship* social choice function, individual and collective pivotality coincide, and allow for a protocol without any contextual privacy violation. We show in Section 6.B.1 that the serial dictatorship (with deterministic lexicographic tie-breaking) is contextually private.[14] It is intuitive that the serial dictatorship does not produce a conflict between collective and individual pivotality—it is impossible to be collectively pivotal without being individually pivotal, since one agent must choose a different object to lead to a different allocation.

Other object assignment choice rules fail to be contextually private because of a conflict between individual and collective pivotality. Consider first the house assignment problem Shapley and Scarf 1974. All agents are initially endowed with an object from C. Denote the initial assignment by an injective function $e \colon N \to C$, where $e(i) \in C$ refers to agent i's initial endowment. For our result it will be irrelevant whether the endowments are private information or known to the designer. We call a choice rule ϕ *individually rational* if for all $i \in N$ and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\phi_i(\boldsymbol{\theta}) \succeq_i e(i).$$

Proposition 6.3. *Assume $|N| \geq 2$. Every protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$ for an individually rational and efficient housing assignment choice rule ϕ produces contextual privacy violations at any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ in which two agents prefer their endowment over the other agent's endowment.*

[14] In fact, serial dictatorships satisfy two even stronger notions of privacy protection than contextual privacy. They satisfy *individual* and *group* contextually privacy, as discussed in Section 6.E.

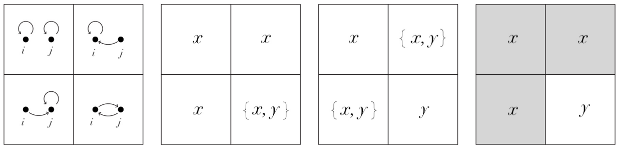


Figure 6.2: Constructing collective but not individual pivotality for house assignment. Type profiles $\{\theta_i, \theta'_i\} \times \{\theta_j, \theta'_j\}$ used in the proof (left, arrows denote whether agent prefers own or other's endowed object); required outcomes for each type profile under individual rationality (mid-left), under efficiency (mid-right), and under both efficiency and individual rationality (right).

Proof. The proof shows that there is collective but not individual pivotality. Consider two agents i and j and two possible preference profiles for each agent. For agent i, consider a type θ_i which contains $e(i) \succ_i e(j)$, and a type θ'_i which contains $e(j) \succ_i e(i)$. For agent j, consider θ_j which contains $e(j) \succ_j e(i)$, and a type θ'_j which contains $e(i) \succ_j e(i)$. Hold fixed all other types $\boldsymbol{\theta}_{-ij}$, to be such that they prefer their own endowment to all other objects, *i.e.* $\boldsymbol{\theta}_{-ij} = (e(k) \succeq_k c \text{ for all } c \in C \setminus \{e(k)\})_{k \in N \setminus \{i,j\}}$.

When the type profile is $(\theta_i, \theta_j, \boldsymbol{\theta}_{-ij})$, the agents both prefer their own endowment to the other's. When the profile is $(\theta'_i, \theta_j, \boldsymbol{\theta}_{-ij})$, or $(\theta_i, \theta'_j, \boldsymbol{\theta}_{-ij})$, they both prefer i's endowment and j's endowment, respectively. When $(\theta'_i, \theta'_j, \boldsymbol{\theta}_{-ij})$, they each prefer the other's endowment to their own. Let x be the outcome in which both agents retain their endowment, *i.e.* $x = (i, e(i)), (j, e(j))$. Let y be the outcome in which each agent gets each other's endowment $y = (i, e(j)), (j, e(i))$. Then, individual rationality makes the requirements shown on the mid-left in Figure 6.2: $\phi(\theta_i, \theta_j, \boldsymbol{\theta}_{-ij}) = (\theta_i, \theta'_j, \boldsymbol{\theta}_{-ij}) = (\theta'_i, \theta_j, \boldsymbol{\theta}_{-ij}) = x$. Meanwhile, efficiency requires $\phi(\theta_i, \theta_j, \boldsymbol{\theta}_{-ij}) = y$ (shown on the mid-right in Figure 6.2). Hence, by Theorem 6.1, every individual elicitation protocol for an individually rational and efficient choice rule produces contextual privacy violations. □

The failure of contextual privacy in the house assignment problem is illuminating. There will be a violation at any type profile $\boldsymbol{\theta}$ in which there is a pair of agents who each prefer their endowment to the other's. Note that in a setting with many agents and many objects, there are many such type profiles $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$. In such cases, the conjunction of efficiency and individual rationality produce an instance where the agents are collectively but not individually pivotal.

In two-sided matching, we see a similar failure mode for contextual privacy and stable outcomes under individual elicitation, but here it is the requirement of stability that produces a conflict between collective and individual pivotality. In two-sided matching, every agent (also called "student") is matched to at most one object (also called "school"), and at most $\kappa(c)$ agents are matched to an object c, for every $c \in C$, for some *capacities* $\kappa(c)$. That is, the set of outcomes is

$$X = \big\{\mu \subseteq N \times C\colon\ \forall i \in N\colon |\{c \in C \mid (i,c) \in \mu\}| \leq 1 \\ \text{and } \forall c \in C |\{i \in N\colon (i,c) \in \mu\}| \leq \kappa(c)\big\}\,.$$

We say there is *no oversupply* if the aggregate capacity equals the number of agents, $\sum_{c \in C} \kappa(c) = n$.

We assume that the objects' preferences over agents are determined by agents' *priority scores*, which are private information of the agents. This matches the *college assignment problem* with standardized test scores (Balinski and Sönmez 1999). More formally, each agent has a vector of *scores* score_c, representing their score at each object $c \in C$. Objects prefer agents with higher scores. Agent i has private information $\theta_i = (\preceq_i, \text{score}_i)$, where $\preceq_i$ is i's preference ranking over schools, and $\text{score}_i \colon C \to \mathbb{R}$ maps objects to scores.

In such school choice settings, a desirable property of choice rules is *stability* (or does not induce *justified envy*). A choice rule ϕ is *stable* or *induces no justified envy* if there is no blocking pair (i, c), $i \in N$, $c \in C$ such that $c \succ_i \phi_i(\boldsymbol{\theta})$ and $\text{score}_i(c) > \text{score}_i(\phi_i(\boldsymbol{\theta}))$.

Proposition 6.4. *Assume $|C| \geq 2$ and $|N| \geq 2$ and that there is no oversupply. Then any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$ for a stable choice rule produces contextual privacy violations.*

This proposition highlights one way in which stability produces contextual privacy violations. Whether an agent has justified envy of another student is a collective feature of both agents' types. A single student's change in score may lead to a new justified claim of another student to their seat in this school, yet no envy. A change in this other student's preference may lead to envy which is not justified. Yet both changes together lead to justified envy, and force collective pivotality in these changes in type.

We can also see, for the school choice problem, how common protocols will produce contextual privacy violations. Consider for example the deferred acceptance protocol, which produces a stable outcome. The designer has to acquire information about tentative assignments that may not be final, and in so doing violates contextual privacy.

Application to Auctions.

Next, we consider how single-item auctions and double auctions produce contextual privacy violations.

Consider a standard private values auction environment in which a single indivisible item is to be allocated to one agent. Agent types are real numbers $\Theta \subseteq \mathbb{R}_+$. The outcomes are given by $x = (q_i, t_i) \in X$, with $q_i \in \{0, 1\}$, $t_i \in \mathbb{R}$, where q_i is agent i's allocation (with $q_i = 1$ meaning that the agent receives the object), and t_i is their payment. Agents have quasilinear preferences defined by

$$u_i((q, t); \theta_i) = \theta_i q_i - t_i, \tag{6.2}$$

We call a single-item auction choice rule *standard* if there is at most one agent $i \in N$ such that $t_i \neq 0$, and $q_i = 1$. We call an auction *efficient* if it maximizes agent welfare, *i.e.*

$$\phi(\boldsymbol{\theta}) \in \arg\max_{(q(\boldsymbol{\theta}), (\boldsymbol{\theta}))} \sum_{i \in N} u_i((q(\boldsymbol{\theta}), t(\boldsymbol{\theta})), \theta_i)$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. That is, the highest valuation agent wins.

The two most widely studied standard auction rules are the first-price and the second-price auction. As this article considers deterministic mechanisms, we consider these rules with deterministic lexicographic tiebreaking. The *first-price auction* is a choice rule $\phi^{\text{FP}}(\boldsymbol{\theta}) = (\phi_1^{\text{FP}}(\boldsymbol{\theta}), \ldots, \phi_n^{\text{FP}}(\boldsymbol{\theta})) = ((q_1, t_1), \ldots, (q_n, t_n))(\boldsymbol{\theta})$, where

$$\phi_i^{\text{FP}}(\boldsymbol{\theta}) = \begin{cases} (1, b_i(\theta_i))) & \text{if } b(\theta_i) = \min \arg\max_{j \in N} b(\theta_j) \\ (0,0) & \text{otherwise.} \end{cases}$$

for some monotonic bid function $b_i \colon \boldsymbol{\theta}_i \longrightarrow \boldsymbol{\theta}$. The *second-price auction* is a choice rule $\phi^{\text{SP}}(\boldsymbol{\theta})$, where

$$\phi_i^{\text{SP}}(\boldsymbol{\theta}) = \begin{cases} (1, \theta_{[2]}) & \text{if } i = \min \arg\max_{j \in N} \theta_j \\ (0,0) & \text{otherwise,} \end{cases}$$

where $\theta_{[2]}$ is the second-highest type in the type profile $\boldsymbol{\theta}$. Both of these auction rules can be computed via a number of different protocols. Commonly studied protocols for the first-price and second-price rules, respectively, include the descending ("Dutch") protocol and the ascending ("English") protocol.

The first-price auction does not at any point lead to an instance of collective pivotality without individual pivotality—only the winner determines the outcome. Indeed, there is a protocol for the first-price auction rule that produces no contextual privacy violations. In particular, the familiar descending or "Dutch" protocol is fully contextually private. We present and discuss this result in Section 6.B.2. The intuition for this result is very similar to the intuition behind the result that the serial dictatorship is contextually private. The designer begins at the top of the type space and asks questions of the form "Is your type above $\tilde{\theta}$?" As soon as one agent answers in the affirmative at some $\tilde{\theta}$, the protocol ends and assigns the object to the agent who responded affirmatively and the price is set at $t = \tilde{\theta}$. The designer thus only elicits information that directly changes the outcome.

Proposition 6.5. *Assume $|\Theta| \geq 3$ and $|N| \geq 3$. Any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$ for the second-price auction choice rule produces contextual privacy violations for $\boldsymbol{\theta}$ at which exactly two agents hold the second-highest type.*

Proof. The proof constructs an instance of collective but not individual pivotality. Consider a type profile with $\theta_{[1]} > \theta_{[4]}$ (or no constraint if $n = 3$), and consider agents i, j that have types in $[\underline{\theta}, \bar{\theta}]$ such that $\theta_{[4]} < \underline{\theta} < \bar{\theta} < \theta_{[1]}$. Consider the product set $\{\underline{\theta}, \bar{\theta}\} \times \{\underline{\theta}, \bar{\theta}\} \times \bigtimes_{k \in N \setminus \{i,j\}} \theta_k$. This corresponds to a square depicted in Figure 6.3.

Let x be the outcome in which the highest type wins ($q_i = 1$ for $\theta_i = \theta_{[1]}$, $q_i = 0$ otherwise) and pays the price $t_i = \overline{\theta}$. Let x' be the outcome under which the highest type wins and pays the price $t_i = \underline{\theta}$. Then, $\phi^{\text{SPA}}(\overline{\theta}, \overline{\theta}, \boldsymbol{\theta}_{-ij}) = \phi^{\text{SPA}}(\underline{\theta}, \bar{\theta}, \boldsymbol{\theta}_{-ij}) = \phi^{\text{SPA}}(\bar{\theta}, \underline{\theta}, \boldsymbol{\theta}_{-ij}) = x$. But, under ϕ^{SPA}, it must be the case that $\phi(\underline{\theta}, \underline{\theta}, \boldsymbol{\theta}_{-ij}) = x'$. Since $x \neq x'$, the is violated, and thus the second-price choice rule is not contextually private under individual elicitation. □

This result is intuitive: when two agents redundantly determine the price, they are collectively pivotal but not individually pivotal. We show in Section 6.A.5 that there are contextual privacy violations even if ties are ruled out (for $|N| \geq 3$ and $|\Theta| \geq 9$).

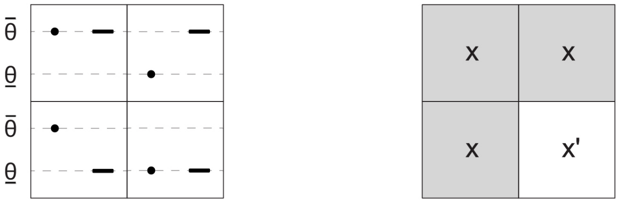


Figure 6.3: Showing collective but not individual pivotality to the second-price auction. Type profiles for agent i and agent j (left, agent j's type is represented by a dot, and agent i's type is represented by a dash); required outcome under the second-price auction rule ϕ^{SPA}.

A similar result holds for standard double auction rules. Suppose m agents are buyers and m agents are sellers, and $n = 2m$. The m sellers are each endowed with one homogeneous, indivisible object. The buyers have unit demand for objects. Formally, agents have initial endowments $e_i \in \{0, 1\}$ where $e_i = 0$ for buyers and $e_i = 1$ for sellers. The preferences are

$$u_i((q, t), \theta) = -e_i \theta_i q_i + (1 - e_i \theta_i) q_i + t_i.$$

A double auction price rule seeks to find a price t that maximizes $\sum_{i \in N} u_i((q, t); \theta_i)$ if buyers with types $\theta_i \geq t$ buy a good at price t, and sellers with value $\theta_i \leq t$ sell their good at price t. Agents with $\theta_i = t$ sell or buy in order to match supply to demand.

Proposition 6.6. *Assume $|\Theta| \geq 4$ and $|N| \geq 4$. Any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$ for an efficient, uniform-price double auction price rule produces contextual privacy violations for $\boldsymbol{\theta}$ where two agents determine the clearing price.*

The intuition for this result is similar to the result for the second-price auction rule. The agents that, together, determine the clearing price are jointly pivotal but not individually so. The proof of this statement requires additional care because the median of an even number of types is not uniquely determined.

Application to Voting and Information Aggregation.

We next turn to a voting environment. The outcome space is ordered, $(X, \leq)$. The preference domain for all agents are the single-peaked preferences with respect to $\leq$.[15]

We consider a commonly studied class of voting rules. Namely, we study *generalized median voting rules*. As shown in Moulin (1980), this class is the class of all anonymous, strategy-proof and Pareto-efficient voting rules, where anonymity means that the outcome cannot depend on the identity of any agent. A generalized median voting rule takes as input submitted peaks of agents' preferences $\theta_1, \theta_2, \ldots \theta_n$ as well as *phantom ballots* $k_1, k_2, \ldots, k_{n-1} \in X \cup \{-\infty, \infty\}$, where for all $x \in X$, $-\infty < x < \infty$. The output is the median of the submitted votes and phantom votes, *i.e.*

$$\phi_{(k_1, k_2, \ldots, k_{n-1})}(\boldsymbol{\theta}) = \text{median}(\theta_1, \theta_2, \ldots, \theta_n, k_1, k_2, \ldots, k_{n-1}).$$

[15] A preference $\preceq$ on $(X, \leq)$ is single-peaked if $x < x' \leq \theta_i \implies x' \succ_i x$ and $x > x' \geq \theta_i \implies x' \succ_i x$.

Proposition 6.7. *Assume $|N| \geq 2$ and $|X| \geq 2$ and the preference domain of single-peaked preferences. Any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$ for a generalized median voting rule that is neither the maximum nor the minimum rule produces contextual privacy violations.*

In such voting rules, classical examples of collective but not individual pivotality arise from voting thresholds, *e.g.*, a qualified majority. Consider a voting rule with at least two phantom ballots k_i on an alternative x that would win under a type profile $\boldsymbol{\theta}$. If there are two agents on the right of this alternative, and *both* of them having a type on the left of x, it would change the outcome—they are *collectively pivotal*—and yet it is still possible that neither of them alone would change the outcome.

6.4 Maximal Contextual Privacy

In the previous section, we focused on how specific properties of the social choice rule lead to contextual privacy violations under individual elicitation technologies. In this section, we take the perspective of a privacy-conscious designer who needs to choose some protocol for a given choice rule. Such a designer must make choices about whose privacy to protect, and at what type profiles. They would like to find protocols that are maximally contextually private—*i.e.* protocols that implement the choice rule, and are maximal elements in the contextual privacy order.

The results of the previous section illuminate a key dimension of the designer's choice—when there is a conflict between collective pivotality and individual pivotality, the designer has a choice of whom to query first, and whose privacy to protect. Those agents who are queried first have their privacy violated. We will see that maximally contextually private protocols protect the privacy of a set of agents i by delaying questions to them as long as possible, so that the designer learns as much as possible about $\boldsymbol{\theta}_{-i}$ before getting to agent i. Note that if everything relevant to the choice rule is revealed about $\boldsymbol{\theta}_{-i}$ before asking anything to agent i, agent i will be asked only what is needed, and will not have a contextual privacy violation.

So, maximally contextually private protocols delay asking questions to protected agents. But how does the designer choose which agents to protect? Maximally contextually private protocols protect different sets of agents and highlight the normative tradeoffs that the designer must consider in choosing a protocol. For example, the two maximally contextually private protocols for the second-price auction rule we derive in Theorem 6.3, the ascending-join and the overdescending-join protocols, protect the winner and the losers, respectively. In some cases, such as the 2017 incentive auction described in Milgrom and Segal (2020), it will be more important to protect the winner, while in other settings, such as in repeated interactions like Google's second-price auctions for advertising, it may be more important to protect the losers.

Theorem 6.3 is the most design-relevant result of the paper, and illustrates the power of the contextual privacy perspective we have developed (the maximally contextually private protocols we derive have not, to our knowledge, been described before). As the theorem's proof illuminates, reasoning about maximally contextually private protocols can be very difficult due to the vast combinatorial space of protocols. In the proof of

Theorem 6.3, we rely on a representation theorem, Theorem 6.2, that puts structure on the space of candidate maximally contextually private protocols. The result shows that for choice rules defined on ordered type spaces, it is without loss, from a contextual privacy perspective, to consider only *bimonotonic* protocols. In a *bimonotonic* protocol, every query is a threshold query, and every agent is asked a sequence of queries that is either monotonically ascending or monotonically descending in the thresholds. This tells us that the two key levers of design, from a contextual privacy perspective, are (i) the initial threshold at which to query agents, and (ii) the order in which to query agents.

We begin in Section 6.4.1 with the representation theorem (Theorem 6.2), and then leverage the representation theorem to derive maximally contextually private protocols for the second-price auction rule (Theorem 6.3) in Section 6.4.2.

6.4.1 A Representation Theorem.

The goal of this subsection is to present our representation theorem (Theorem 6.2) for contextual privacy equivalence classes under the restriction to individual elicitation protocols. The theorem shows that a privacy-concerned designer can, without loss, restrict attention to *bimonotonic* protocols under the restriction to individual elicitation protocols $\mathfrak{S}_{\text{IE}}$. This reduces the designer's levers of design: for each agent they choose an initial threshold, and the order in which to query agents.

First, we formally define bimonotonicity. We will assume that message spaces M have at least two elements, and hence protocols can compute some non-constant social choice functions. In every query of a bimonotonic protocol, an agent sends one of two messages m, m' depending on whether their type is larger than a *threshold* $\tilde{\theta}$,

$$\sigma_i(v, \theta_i) = \begin{cases} m & \text{if } \theta_i > \tilde{\theta} \\ m' & \text{if } \theta_i \leq \tilde{\theta}. \end{cases}$$

That is, a query in a *threshold protocol* asks an agent for whether their type falls strictly above some value $\tilde{\theta} \in \Theta$. A bimonotonic protocol is a protocol in which all queries are threshold queries, and the thresholds of threshold queries to a single agent i form an increasing or decreasing interval in the type space Θ.

Definition 6.4 (Bimonotonic Protocol)**.** A threshold protocol $(P, \boldsymbol{\sigma})$ is *bimonotonic* if for any path in P and for all agents $i \in N$, the sequence of thresholds $\tilde{\theta}$ presented to agent i form an increasing or decreasing interval in the type space Θ.

The representation theorem will show that for many choice rules of interest, any $\mathfrak{S}_{\text{IE}}$-protocol is contextual privacy equivalent to a bimonotonic protocol. The condition for this reduction is that the choice rule of interest must be injective over exactly one interval inside of the type space, a property which we call *interval pivotality.*

Definition 6.5 (Interval Pivotality)**.** A social choice function ϕ defined on an ordered type space Θ exhibits *interval pivotality* if for all $\boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}_{-i}$ there are elements $\underline{\theta} \in \Theta$ and $\overline{\theta} \in \Theta$ such that

$$\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta'_i, \boldsymbol{\theta}_{-i}) \iff (\theta_i, \theta'_i \leq \underline{\theta} \text{ or } \theta_i, \theta'_i \geq \overline{\theta}).$$

This property of choice rules states that, holding fixed some profile of other agents' types $\boldsymbol{\theta}_{-i}$, the choice rule $\phi(\theta_i, \boldsymbol{\theta}_{-i})$ is constant in θ_i if and only if θ_i is outside some interval $[\underline{\theta}, \overline{\theta}] \in \Theta$. So, inside the interval, i is "pivotal" in that her report changes the outcome. Note in this definition that the interval defined by $[\underline{\theta}, \overline{\theta}]$ may depend on the profile of other agents' types, $\boldsymbol{\theta}_{-i}$.[16]

Many social choice rules of interest are interval pivotal. For example, all auction and voting social choice functions considered in Section 6.3.2 are interval pivotal. Note, however, that interval pivotality only applies to choice rules defined on ordered type spaces, so assignment and matching rules cannot be interval pivotal. And not all choice rules on ordered type spaces are interval pivotal. For example, the third price auction is not interval pivotal—there are two disjoint intervals of θ_i over which the choice rule $\phi(\theta_i, \boldsymbol{\theta}_{-i})$ is changing (when θ_i is setting the price, and when θ_i is determining the winner).

Now that we have defined bimonotonicity and interval pivotality, we can present the main result of this subsection, and indeed a central theoretical insight of the paper. This representation theorem simplifies the designer's search for privacy-preserving protocols when choosing among individual elicitation protocols for an interval pivotal choice rule: the designer need only consider bimonotonic protocols, because every $\mathfrak{S}_{\text{IE}}$-protocol is contextual privacy equivalent to a bimonotonic protocol.

Theorem 6.2. *Let $(P, \boldsymbol{\sigma})$ be a $\mathfrak{S}_{\text{IE}}$-protocol for ϕ, where ϕ exhibits interval pivotality. There is a bimonotonic protocol $(P', \boldsymbol{\sigma}')$ that has the same set of contextual privacy violations as $(P, \boldsymbol{\sigma})$,* i.e. $\Gamma(P, \boldsymbol{\sigma}, \phi) = \Gamma(P', \boldsymbol{\sigma}', \phi)$.

In other words, every contextual privacy equivalence class has a bimonotonic representative. The proof of Theorem 6.2 proceeds by considering an arbitrary protocol and applying two modifications to it. These modifications transform the arbitrary protocol into a bimonotonic one, while preserving the set of contextual privacy violations. The first transformation "fills in" threshold queries for any threshold that lies between any types that are distinguished. Second, a "scrubbing" operation keeps only a monotonic sequence of threshold queries. It is not difficult to see that the result of these operations will be a bimonotonic protocol. What is less obvious is why contextual privacy violations do not change. Interval pivotality is crucial here—it implies that when filling in gaps between threshold queries, the protocol is either filing in inside of the "pivotal interval" $[\underline{\theta}, \overline{\theta}]$ (in which case no violations are introduced because the agent is changing the outcome) or outside the pivotal interval (in which case there was already a violation).

Several common protocols are bimonotonic: The ascending protocol for the second-price auction, the descending protocol for the first-price auction, and the overdescending protocol for the second-price auction. For voting rules, one-sided protocols are bimonotonic: First, query all agents whether they are below the right-most alternative, then query the next-to-rightmost alternative, and so forth.

The fact that every equivalence class has a bimonotonic representative emphasizes that the decisions a designer makes about contextual privacy amount to: (i) the threshold of the initial query to each agent, and (ii) the order in which the designer asks agents

[16] We suppress subscripts to simplify notation. Formally, the interval is $[\underline{\theta}_{(i,\boldsymbol{\theta}_{-i})}, \overline{\theta}_{(i,\boldsymbol{\theta}_{-i})}]$.

these threshold queries. So, in searching for maximally contextually private protocols in particular, it is without loss to focus on these two dimensions. We leverage this theorem in the next subsection.

6.4.2 Two Maximally Contextually Private Protocols for the Second-Price Auction Rule.

Next, we use the representation theorem to guide design: we find maximally contextual private protocols for the second-price auction rule. We will derive two novel protocols, *ascending-join* and *overdescending-join* protocols, that are maximally contextually private.

We begin by formalizing the idea of a protection set, *i.e.* some set of agents and type profiles $A \subseteq N \times \boldsymbol{\Theta}$ that do not have privacy violations under P.

Definition 6.6. We say that a protocol P *protects* $A \subseteq N \times \boldsymbol{\Theta}$ if

$$\Gamma(P, \phi) \cap A = \emptyset.$$

Denote the set of $\mathfrak{S}$-protocols that protect A by $\mathcal{P}^A_{\mathfrak{S}}$.

Protection sets are helpful constructs for communicating the contextual privacy properties of particular protocols. For example, Milgrom and Segal (2020) study protocols that preserve *unconditional winner privacy* which is analogous, in our formalism, to protocols that protect only the contextual privacy of winners. We will use protection sets that also include losers.

The following observation helps us in characterizing maximally contextually private protocols. It is direct from the definition of the contextual privacy order defined by inclusion of the set of contextual privacy violations. It says simply that if a protocol is maximally contextually private among a subset of $\mathfrak{S}$-protocols that protect a subset A, *i.e.* among $(\mathcal{P}^A_{\mathfrak{S}}, \preceq_\phi)$, then that protocol is also maximally contextually private among all $\mathfrak{S}$-protocols.

Lemma 6.1. *Assume that P is maximal in $(\mathcal{P}^A_{\mathfrak{S}}, \preceq_\phi)$. Then, it is maximal in $(\mathcal{P}_{\mathfrak{S}}, \preceq_\phi)$.*

We will use Lemma 6.1 to show maximality of protocols for ϕ^{SPA}. We define specific protection sets A and characterize maximally contextually private protocols within $\mathcal{P}^A_{\mathfrak{S}}$, as it is easier to prove maximal contextual privacy within $\mathcal{P}^A_{\mathfrak{S}}$ than within $\mathcal{P}_{\mathfrak{S}}$ directly.

The protocols we show to be maximally contextually private protect specific agent-type profile pairs $(i, \boldsymbol{\theta})$. In the definition of these sets, tiebreaking requires extra care. We call an agent a *winner* if she is the highest (in lexocographic order) priority, highest-type agent (as usual). We say that an agent *determines the price* if she is the highest priority, second-highest type agent (or the second-highest priority, highest type agent in the case of a tie). If the agent neither determines the price nor is a winner, then she is a *loser*. These definitions give rise to the winner and loser protection sets, W and L .

$$\begin{aligned} \mathrm{W} &= \{(i, \boldsymbol{\theta}) \mid i \text{ is the winner at } \boldsymbol{\theta}\} \\ \mathrm{L} &= \{(i, \boldsymbol{\theta}) \mid i \text{ is neither the winner nor determines the price at } \boldsymbol{\theta}\}. \end{aligned}$$

We also define subsets of the winner and loser protection sets. Let $d(\boldsymbol{\theta})$ be the index of the agent who determines the price at type profile $\boldsymbol{\theta}$. We define the protection sets LPL and HPW as follows:

$$\begin{aligned} \mathrm{LPL} &= \mathrm{L} \cap \{(i, \boldsymbol{\theta}) \mid i > d(\boldsymbol{\theta})\} \\ \mathrm{HPW} &= \mathrm{W} \cap \{(i, \boldsymbol{\theta}) \mid i < d(\boldsymbol{\theta})\}. \end{aligned}$$

Low-priority losers are losers whose index i is greater than the index of the agent with the second-highest type. Note that agents with high indices are "low priority" because of lexicographic tie-breaking. High-priority winners are winners whose index is lower than the index of (or have higher priority than) the agent with the second-highest type.

The next theorem states that the maximally contextually private protocols are simple. that there is a unique protocol for the second-price auction rule that protects the winner and low-priority losers, and a unique protocol for the second-price auction rule that protects the losers and high-priority winners. These two protocols are maximally contextually private.

Theorem 6.3. *For the second-price auction rule ϕ^{SPA}, there is:*

(a) a unique minimal set of contextual privacy violations for W- *and* LPL*-protecting $\mathfrak{S}_{\mathrm{IE}}$-protocols,*
(b) a unique minimal set of contextual privacy violations for L- *and* HPW*-protecting $\mathfrak{S}_{\mathrm{IE}}$-protocols.*

Two bimonotonic protocols attain these sets of contextual privacy violations, which we call the ascending-join protocol *and the* overdescending-join protocol*, respectively. Both of these protocols are maximally contextually private among $\mathfrak{S}_{\mathrm{IE}}$-protocols for ϕ^{SPA}.*

The basic idea of the result is that the designer delays asking questions to the agents in the protection set as much as possible, so that on type profiles where it is possible to avoid violating their privacy, the designer will avoid violating their privacy. The two protocols we derive are variants of more common protocols. The ascending-join protocol is related to the standard ascending or "English" auction protocol that begins with queries at the bottom of the type space and stops when only one agent remains, and the overdescending-join protocol is related to the overdescending protocol (see Harstad (2018)) which begins at the top of the type profile, and descends until the second-highest value is found.

The proof begins with the restriction to bimonotonic protocols via Theorem 6.2. Recall that the restriction to bimonotonic protocols reduces the designer's problem to choices around (i) the order in which to query agents and (ii) the initial threshold query to pose to each agent. We show that at every point in the protocol, there is exactly one option for (i) a next agent to ask, and (ii) the threshold strategy for this agent that maintains the privacy of the protection sets in the hypothesis.

We give further intuition for the ascending-join protocol here, and refer readers to the appendix for further understanding of the overdescending-join protocol. In the proof of Theorem 6.3, we show that any $\mathrm{W} \cup \mathrm{LPL}$ protecting protocol can be understood in "rounds". We show that after each such round but the last, the designer's knowledge is a *protective information state* $\boldsymbol{\Theta}_v$. We go on to show that from protective information states, W ∪ LPL protection demands a unique sequence of "non-redundant" queries, which we show to imply the result.

Here we give further intuition for the information state reached at the end of a round. Any W- and LPL-protective information state has a running price $\tilde{\theta}$ and an index of the most recently "joined" agent l such that: (i) nothing is known about agents with index higher than l (*idle agents*); (ii) there are exactly two agents for whom it is exactly known that their type is above $\tilde{\theta}$ (*active agents*); and (iii) all other agents are known to have a type below $\tilde{\theta}$ (*dropped-out agents*). These requirements are illustrated in Figure 6.4. With this definition of a protective information state, we then show that at every possible stage of the protocol, there is exactly one query that the designer can make that leads into either another protective information state or termination.

The ascending-join protocol can roughly be understood as follows. The designer wants to maintain exactly two active agents at any point in time, until termination. The protocol begins with asking the second-highest priority agent (agent with index 2) the lowest threshold query. When the designer gets an affirmative answer for some threshold query, they next ask an agent with higher priority than the queried agent the same threshold query. When the designer gets a negative answer, they ask an agent of lower priority next.

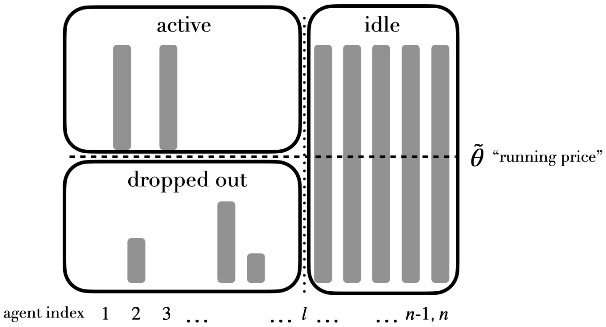

active
idle
$\tilde{\theta}$ "running price"
dropped out
agent index 1 2 3 … … l … … n-1, n

Figure 6.4: W- and LPL-Protective Information State.

To illustrate what the ascending-join protocol actually looks like in practice, we turn to a simple example with four agents.

Example 6.1 (Ascending-Join Protocol for the Second-Price Auction Rule). Consider a set of four agents $|N| = 4$ and a type space of the integers from 1 and 10, *i.e.* $\Theta = \{1, 2, \ldots, 10\}$. Suppose the true type profile is $\boldsymbol{\theta} = (4, 3, 8, 2)$.

For this type profile, the winner is agent 3, who has value $\theta_3 = 8$. There is a single low-priority loser, agent 4 who has value $\theta_4 = 2$. The ascending-join protocol for this type profile is spelled out in Table 6.3, and illustrated (with a comparison to the ascending protocol) in Figure 6.5.

Note in the ascending-join protocol that at every stage except the terminal stage, the designer is in or leading into a *protective information state*: there is a running price $\tilde{\theta}$, there are two active agents who are known to have types above $\tilde{\theta}$ (the agents who say yes for each threshold), any agents who have dropped out have indices to the right of an active agent (*e.g.*, when agent 2 drops out at running price $\tilde{\theta} = 3$, she is to the right of agent 1, who is active) and there is a set of idle agents about whom nothing is known (agent 4 is

$\tilde{\theta} = 1$	Query agent 2, "Is your type above 1?" Agent 2 answers **yes**.
	Query agent 1, "Is your type above 1?" Agent 1 answers **yes**.
$\tilde{\theta} = 2$	Query agent 2, "Is your type above 2?" Agent 2 answers **yes**.
	Query agent 1, "Is your type above 2?" Agent 1 answers **yes**.
$\tilde{\theta} = 3$	Query agent 2, "Is your type above 3?" Agent 2 answers no.
	Query agent 3, "Is your type above 3?" Agent 3 answers **yes**.
	Query agent 1, "Is your type above 3?" Agent 1 answers **yes**.
$\tilde{\theta} = 4$	Query agent 3, "Is your type above 4?" Agent 3 answers **yes**.
	Query agent 1, "Is your type above 4?" Agent 1 answers no.
	Query agent 4, "Is your type above 4?" Agent 4 answers no.
The protocol terminates, allocates the object to agent 2 at price $\tilde{\theta} = 4$.	

Table 6.3: Example: Ascending-Join Protocol for $\boldsymbol{\theta} = (4, 3, 8, 2)$.

idle until the final stage). The winner (agent 2) and the lowest-priority loser (agent 4) do not have contextual privacy violations. Compare this to the ascending protocol, which additionally violates the contextual privacy of agent 4, as shown in Figure 6.5.

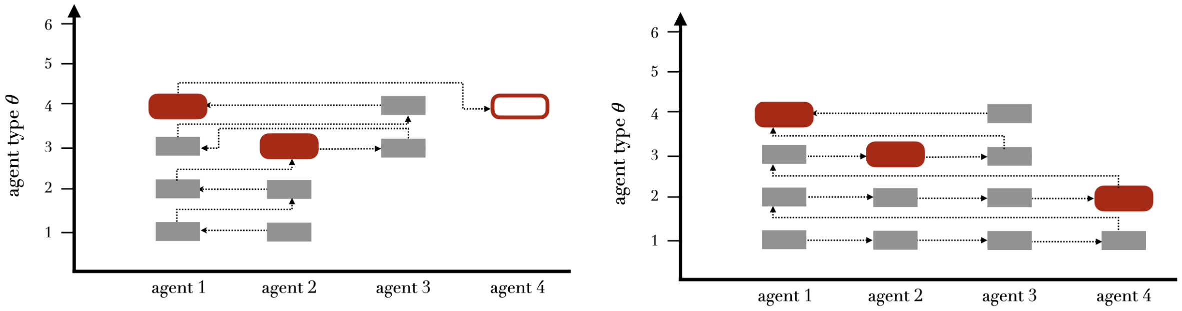


Figure 6.5: Example—Ascending-Join (left) vs. Ascending (right) Protocol for $\boldsymbol{\theta} = (4, 3, 8, 2)$. Queries to an agent are represented as rectangles, positioned at the corresponding threshold $\tilde{\theta}$ (gray solid rectangles represent affirmative responses, red rounded rectangles represent negative responses and a contextual privacy violation, red unfilled rectangles represent negative responses without a contextual privacy violation). The path of the dotted line represents the order in which queries are asked. In both protocols, the winner (agent 3) does not have their privacy violated. In the ascending-join protocol, the low-priority loser (agent 4) also does not have their privacy violated. Theorem 6.3 shows that the ascending-join protocol is $\mathfrak{S}_{\mathrm{IE}}$-maximally contextually private.

The constructive proof strategy of Theorem 6.3 can be understood through this example. The basic idea is that all queries must lead into a protective information state or termination, and there is a unique sequence of queries that do so. In this sequence, we show in a lemma that it must never be the case that a query to a particular agent could land them in the protection set *regardless of their answer*. That is, no agent i can never be asked a query at node v that could, for some $\boldsymbol{\theta}_{-i} \in \text{children}(\boldsymbol{\Theta}_v)$, lead to $i \in W \cup \text{LPL}$ regardless of i's answer to the query at v.

This lemma is very powerful. To see how we use it to construct the protocol, consider the initial query in the example. We know from bimonotonicity that we must determine the initial query's threshold, and the agent to ask it to. First off, we know that the initial query must have a threshold $\tilde{\theta} = 1$, asking about the lowest type, because otherwise the winner's privacy could be violated. The more interesting part is: to whom does the designer ask the first query? The initial query must go to agent 2. To see this, note that if the designer asked agent 1 whether her type is above $\tilde{\theta} = 1$, there are cases in which she ends up a winner regardless of her answer. If she answers "no" (*i.e.* she has the lowest possible value) and it turns out that all other agents have the lowest possible value, she wins the object, and at the same time, if she answers "yes" (*i.e.* she has a value above the lowest), and everyone else has the lowest value, then again agent 1 wins the object. So, regardless of agent 1's answer, she could end up in the protection set, and so we cannot ask the initial query to agent 1. If the initial query with threshold $\tilde{\theta} = 1$ went to agents 3 or 4, note also that they could end up in the protection set regardless of their answers. If agent 3 or 4 answered "yes", they could be a low-priority loser, and if agent 3 or 4 answered "no" they could be a low-priority loser. So, the only remaining option for the initial query is to query agent 2. For agent 2, a positive answer could lead them into the protection set (they could end up the winner) but a negative answer could not lead them into the protection set (they could be a loser, but they could not be a low-priority loser).

The rest of the constructive proof proceeds in a similar manner. In general, the designer wants to keep two active agents at a time. Starting with the initial query to agent 2, whenever the designer gets an affirmative response, they "move left" asking an agent to the left the same threshold. When the designer hits a negative response, they "move right", *i.e.* another agent *joins* the protocol from the right. For instance, in the example, when agent 2 drops out, the designer "moves right" to agent 3 at the same threshold.

The overdescending-join protocol is constructed in an analogous manner, beginning at the top of the type space, *i.e.* with an initial threshold query of $\tilde{\theta} = \max \Theta$, and asks a descending sequence of threshold queries. As the principles of the construction are similar, we relegate its discussion to the appendix.

We highlight that different maximally contextually private protocols protect different agents. In particular, the ascending-join protocol is not preferable, from a privacy standpoint alone, to the overdescending-join protocol. The two protocols violate the privacy of disjoint sets of agent-type profiles.

Proposition 6.8. *The ascending-join protocol and the overdescending-join protocol are incomparable in the contextual privacy order.*

This result is a consequence of the fact that the ascending-join protocol, while protecting *low priority* losers, does violate the privacy of some losers (*e.g.*, agent 3 in the example). The overdescending-join protocol, while protecting *high priority* winners, does violate the privacy of some winners (*e.g.*, agent 2 in the example). The designer must make a normative decision about whose privacy is more important to protect—for example, in the incentive auction design described in Milgrom and Segal (2020), it is more important to protect winners.

Are the ascending-join and overdescending-join protocols the uniquely maximally contextually private protocols for the second-price auction rule? No, they are not. To show the existence of other maximally contextually private protocols, we need only note that there are protocols that are incomparable in the contextual privacy order to these two. In Section 6.C, we discuss a particular protocol which we call the *guessing protocol* that is incomparable to both the ascending-join and overdescending-join protocols.

6.5 Contextual Privacy and Incentives

In this section, we briefly discuss the strategic aspects of the messaging game induced by an individual elicitation protocols. We formally define two dynamic implementation notions in Section 6.5.1: implementation in dominant and obviously dominant strategies.

In Section 6.5.2 we consider the incentive properties of the maximally contextually private protocols discussed in Section 6.4. We first observe that the ascending-join protocol is implementable in obviously dominant strategies. This result is a direct consequence of Li (2017)'s characterization of obvious dominance. Then, we show that the overdescending-join protocol is implementable in dominant strategies. This result strengthens and formalizes an observation in Harstad (2018) that the overdescending protocol is strategyproof.

6.5.1 Implementation in (Obviously) Dominant Strategies.

Recall that agents submit messages from a set M according to strategies $\sigma_i \colon \Theta \times V \to M$. The state of the protocol is *public*, so that their strategy takes into account information known at node $v \in V$, *i.e.* $\mathbf{\Theta}_v$.

Implementation in dominant strategies requires that at every node v, for every agent i and any opponent message profile $\boldsymbol{b}_{-i}$, sending a message according to the strategy $\sigma_i(\theta_i, v) \in M$ yields a higher utility than any other message $m_{iv} \in M$. To distinguish actions of agent i at node v from actions of agent i at other nodes, we write $\sigma_{i-v} \colon \Theta \to (V \to M)$ for the partial strategy of agent i at nodes other than v.

Definition 6.7. A strategy σ_i in protocol P is *dominant* for agent i if for all θ_i, at all nodes v and possible opponent message profiles $\boldsymbol{b}_{-i}$,

$$u_i\left(P(\sigma_i(\theta_i, v), \sigma_{i-v}(\theta_i), \boldsymbol{b}_{-i}); \theta_i\right) \geq u_i\left(P(m_{iv}, \sigma_{i-v}(\theta_i), \boldsymbol{b}_{-i}); \theta_i\right). \tag{6.3}$$

for any deviation $m_{iv} \in M$. We say that a protocol P for ϕ *implements* ϕ in *dominant strategies* if there exists a dominant strategy σ_i for each agent i such that $P(\boldsymbol{\sigma}(\boldsymbol{\theta})) = \phi(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbf{\Theta}$.

Implementation in dominant strategies demands that at any node and under any play by other agents or at other nodes, it must be optimal for an agent to follow their strategy $\sigma_i(\theta_i, v)$. Note that our notion of dominance assumes that the agent continues to follow her strategies at other nodes σ_{-iv}.

Implementation in obviously dominant strategies requires that the worst outcome following the messaging strategy is better than the best outcome from a deviation, assuming that the deviating agent follows her strategy in future nodes.

Definition 6.8. A strategy σ_i in protocol P is *obviously dominant* for agent i if for all θ_i, at all nodes v,

$$\min_{\boldsymbol{b}_{-i}} u_i \left(P(\sigma_i(\theta_i, v), \sigma_{i-v}(\theta_i), \boldsymbol{b}_{-i}); \theta_i\right) \geq \max_{\boldsymbol{b}_{-i}} u_i \left(P(m_{iv}, \sigma_{i-v}(\theta_i), \boldsymbol{b}_{-i}); \theta_i)\right),$$

for all potential deviations $m_{iv} \in M$. We say that a protocol P for ϕ *implements* ϕ in *obviously dominant strategies* if there exists an obviously dominant strategy σ_i for each agent i such that $P(\boldsymbol{\sigma}(\boldsymbol{\theta})) = \phi(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

In an obviously dominant strategy profile, it has to be the case that for all agents i, the worst outcome that can result from following the strategy σ_i, is better than the best outcome that can result from some other strategy σ_i', regardless of what other agents do in the future.

6.5.2 Dynamic Incentives of Maximally Contextually Private Protocols.

We now show that the two maximally contextually private designs highlighted in Section 6.4 also have good incentive properties. We begin with the ascending-join protocol.

Proposition 6.9. *The ascending-join protocol implements the second-price auction choice rule ϕ^{SPA} in obviously dominant strategies.*

THe second part of Proposition 6.9 is a direct result of known properties of *personal-clock auction*, which are a class that contains the ascending-join protocol. Therefore, we can rely on the characterization in Li (2017, Theorem 3), which states that every personal-clock auction protocol has an equilibrium in obviously dominant strategies.

We also show that the second protocol shown to be maximally contextually private in Theorem 6.3 has a dynamic implementation, albeit with a weaker incentive guarantee.

Proposition 6.10. *The overdescending-join protocol implements the choice rule ϕ^{SPA} in dominant strategies.*

To see that this statement holds, observe that there are two phases of the overdescending-join protocol: a phase before the winner is found, and a phase after the winner is found. Before the winner is found, for any fixed opponent message profile $\boldsymbol{b}_{-i}$, the price of the good is not influenced by the action of the agent i. The agent i only influences whether they win the good or not—they have an incentive to send a message that they are above a threshold as long as they prefer winning over losing, which is the case if and only if the current price is at or below their type. In the second phase, once the winner has been found, all agents are indifferent between all messages. So, this set of strategies is clearly weakly dominant.

In addition to the ascending-join and overdescending-join protocols, the serial dictatorship also has good privacy properties and incentive guarantees. In particular, we show in Section 6.B that the serial dictatorship protocol is contextually private (*i.e.* it produces no privacy violations) and that the messaging strategies are obviously dominant.

We view further study of incentives and contextual privacy as a direction for future work. The intention of this section was primarily to show that our formalism is capacious enough to discuss incentives, and to begin to show how our results connect to the literature on dynamic implementation.

6.6 Related Literature

This paper brings privacy considerations into extensive-form mechanism design. We discuss here our connection to extensive-form mechanism design, as well as other literature on designing for privacy in computer science and cryptography.

Our restriction to individual elicitation protocols coincides with the extensive-form messaging game used to define and study obvious strategyproofness (Li 2017) and credibility (Akbarpour and Li 2020). Credibility shares a motivation with contextual privacy—both criteria have to do with the potential for the designer to somehow abuse their communication channel with the agent. Credibility, which requires incentive compatibility for the auctioneer, sometimes coincides with the diagnoses of contextual privacy and sometimes not.[17] Under individual elicitation protocols, contextual privacy can be understood as a form of privacy that is easier for participants to understand, just as obvious strategyproofness is a form of strategyproofness that is easier for participants to understand. Many papers study variants of obvious strategyproofness and their compatibility with other axioms and computational properties (Bade and Gonczarowski 2017; Ashlagi and Gonczarowski 2018; Mackenzie 2020; Golowich and Li 2022; Yiqiu Chen and Westkamp 2022; Pycia and Troyan 2023).

Other considerations related to privacy and trust have been incorporated into mechanism design and market design, in both static and dynamic models. Though it is not the focus of their paper, Mackenzie and Zhou (2022) discuss how the dynamic *menu mechanisms* they define (in which at each history the agent chooses from a menu of possible outcomes) protect privacy compared to direct mechanisms. Grigoryan and Möller (2023) and Woodward (2020) define two different but related notions of the *auditability* of different mechanisms, based on the amount of information that would be required to determine whether the outcome of a mechanism had been correctly computed, and Hakimov and Raghavan (2020) shows how providing feedback to participants can help to verify the mechanism. Others study how the disclosure of past trades affects future trades (Dworczak 2020; Ollar, Rostek, and Yoon 2017) and how the desire for privacy may stem from a desire to avoid price discrimination (Ichihashi 2020; Ali, Lewis, and Vasserman 2022). Canetti, Fiat, and Gonczarowski (2023) considers the privacy of the designer as opposed to our focus on the privacy of agents, and investigates the use of zero-knowledge proofs to prove properties of the mechanism without revealing the designer's objectives. Several papers have incorporated measures of "privacy loss" as constraints on mechanism design (Eilat, Eliaz, and Mu 2021; D. Liu and Bagh 2020), where privacy loss is defined as some measure (*e.g.*, Shannon entropy, Kullback-Leibler divergence) of information revelation. These measure-based criteria treat all data as equal. Contextual privacy, unlike these measure-based criteria, is not about *how much* information is revealed, and is also not just about *whether* information is revealed, but rather it is about *how* the information that is revealed is *used*.

There are two important precursors to contextual privacy in the theory of decentralized computation: unconditional full privacy and perfect implementation. Contextual

[17] The descending or *Dutch* protocol of the first-price auction is both contextually private and credible, but the ascending protocol of the second-price auction is credible but not contextually private.

privacy under the individual elicitation technology parallels the concept of *unconditional full privacy* for decentralized protocols (Chor and E. Kushilevitz 1989; Brandt and Sandholm 2005; Brandt and Sandholm 2008). Unconditional full privacy requires that the only information revealed through a decentralized protocol is the information contained in the outcome—this notion of privacy is *unconditional* in that it does not condition on the presence of a mediator or on computational hardness assumptions. It has been applied to an auction domain (Brandt and Sandholm 2008), and a voting domain (Brandt and Sandholm 2005), stressing impossibility results. Our definition of contextual privacy brings unconditional full privacy into a framework amenable to economic design and extends it in several ways, highlighting that design for privacy involves tradeoffs regarding whose privacy to protect. Milgrom and Segal (2020)'s concept of *unconditional winner privacy* is similar to contextual privacy in that it brings unconditional full privacy into centralized mechanism design: unconditional winner privacy is unconditional full privacy in a centralized mechanism, for the winner only. We generalize the Milgrom and Segal (2020)'s concept to objectives beyond protection of the winner's privacy. The notion of *perfect implementation* (S. Izmalkov, S. Micali, and M. Lepinski 2005; Sergei Izmalkov, Matt Lepinski, and Silvio Micali 2011) seeks implementations that do not rely on trusted mediators, but rather rely on simple technologies that enable verification of what was learned—like sealed envelopes. The construction allows for particular elicitation technologies that allow, *e.g.*, envelopes-inside-envelopes. Although our general set-up can accommodate such technologies, we focus on more minimal assumptions about the technology available to the designer (*e.g.*, the individual elicitation technology), and offer a privacy order rather than an implementation concept.

Beyond unconditional full privacy and perfect implementation lies an extensive literature on privacy preserving protocols for auctions and allocation. The literature on cryptographic protocols for auctions, going back to Nurmi and Salomaa (1993) and Franklin and Reiter (1996) is too vast to summarize here—the main point is that there are many cryptographic protocols that do not reveal *any* private information to a designer. Such protocols allow participants to jointly compute the outcome without relying on any trusted third party while usually relying on computational hardness assumptions. Compared to this literature, contextual privacy makes explicit the social and technological environments in which many designers operate: when arbitrary cryptographic protocols are not available, we need some other privacy desideratum to guide design.

An influential privacy desideratum is differential privacy (Dwork et al. 2006). Contextual privacy sharply diverges from interpretations of differential privacy in mechanism design contexts. Differential privacy was originally defined for database management. For a survey of its incorporation into mechanism design, including works that incorporate privacy concerns in agent utility functions (Nissim, Orlandi, and Smorodinsky 2012; A. Roth and Schoenebeck 2012; Ligett and A. Roth 2012; Ghosh and A. Roth 2015), see Pai and A. Roth (2013). Differential privacy, as adapted for mechanism design contexts, says that the information revealed about a single agent should have a *negligible* effect on the outcome. Contextual privacy, however, says that subsets of the agent's information should be revealed if it has *some* effect on the outcome. Where contextual privacy justifies information revelation through its relevance to a social outcome, differential privacy explicitly restricts the sensitivity of an outcome to revealed information.

More broadly, our paper connects to a growing theoretical and empirical literature on digital privacy in economics. While we cannot do justice to this literature here, Goldfarb and Que (2023) and Acquisti, C. Taylor, and Wagman (2016) offer valuable surveys of it. One preoccupation of this literature is the appearance of a digital "privacy paradox", wherein consumers' stated preferences about privacy (that it matters) do not align with their revealed preferences (which suggest that it doesn't matter). Some explanations of the privacy paradox point out that data externalities may complicate individual incentives to pursue privacy-preserving actions (Ichihashi 2021; Acemoglu et al. 2022; Bergemann, Bonatti, and Gan 2022). The data externalities argument helps to motivate why a social planner would want to design for privacy (using a criterion like contextual privacy) even when operating in a setting where agents don't appear to have privacy concerns. Recent empirical work quantifies preferences for privacy and emphasizes that they are context-dependent and changing over time, suggesting that what appears to be a privacy paradox may simply be a reflection of the highly context-sensitive nature of preferences for privacy (Goldfarb and Tucker 2012; Lin 2022; Tang 2019). This evidence suggests that it is important to have a normative notion of privacy that is attentive to not just *how much* information is revealed but also *how the information revealed is used*—this is what contextual privacy captures.

6.7 Conclusion

An agent's contextual privacy is violated if the designer learns superfluous information about her. It's not always possible for the designer to protect every agent's contextual privacy while still eliciting enough information to compute the outcome of a choice rule. So, the designer needs to make a deliberate decision about whose privacy is most important to protect. In auctions, for example, this means prioritizing either winner or loser privacy. It might be, as is the case for the English protocol for the second-price auction rule, that a commonly-used protocol's privacy can be strictly improved. These improvements come from delaying the involvement of some agents.

Such a delay can be interpreted as a general principle for the design of private systems, such as those allocating personalized experiences: If data from a subset of agents is sufficient to estimate all that is needed to determine outcomes independently for each agents, then there are no contextual privacy violations for the remaining agents: Indeed, it is possible to let the agents send messages determining the *correct* outcome for them.

The contextual privacy order we defined leads to robust comparisons independent of the designer's prior knowledge. An alternative approach, not captured by our analysis, is to specify a cardinal objective for the designer that incorporates a loss from contextual privacy violations. Such a designer can compute an optimal protocol taking into account their privacy objectives. Consider, for example, a case in which all privacy violations incur the same loss in social welfare. In such a case, the designer could use the prior on type profiles to find an optimal protocol from a privacy perspective, weighing the losses from (many) losers against (few) winners.

A remaining concern, for the privacy-conscious designer, is about how contextual privacy interacts with dynamic incentives. While the serial dictatorship and the ascending-join and overdescending-join protocols have good dynamic incentives, it is not clear which properties of these privacy-preserving protocols lead to such incentive guarantees. Absent more structure on the choice rule—such as a strengthening of interval pivotality—incentive properties of privacy-preserving protocols will depend on details of the environment.

6.A Proofs

6.A.1 Proof of Proposition 6.1.

Proposition 6.1. *An $\mathfrak{S}$-computable choice rule ϕ is $\mathfrak{S}$-contextually private if and only if there does not exist a subset of type profiles $\hat{\boldsymbol{\Theta}} \subseteq \boldsymbol{\Theta}$ such that*

(i) $\phi|_{\hat{\boldsymbol{\Theta}}}$ is non-constant, and

(ii) for every partition $\mathcal{S} \in \mathfrak{S}^*$ such that none of the partition cells $S \in \mathcal{S}$ contains $\hat{\boldsymbol{\Theta}}$, there are type profiles $(\theta_i, \boldsymbol{\theta}_{-i}), (\theta'_i, \boldsymbol{\theta}_{-i}) \in \hat{\boldsymbol{\Theta}}$ that lie in different partition cells of $\mathcal{S}$.

Here, $\mathfrak{S}^* \subseteq 2^{2^{\boldsymbol{\Theta}}}$ is the set of revealed partitions. *Define $\mathcal{S}^* \in \mathfrak{S}^*$ if there is a partition $\mathcal{S} \in \mathfrak{S}$ and functions $f_1, f_2, \dots, f_n \colon \Theta \to M$ such that every $S^* \in \mathcal{S}^*$ is the preimage under $(f_1, f_2, \dots, f_n)\colon \boldsymbol{\Theta} \to \mathbf{M}$ of some $S \in \mathcal{S}$, $S^* = (f_1, f_2, \dots, f_n)^{-1}(S)$.*

In addition, if there is such a $\hat{\boldsymbol{\Theta}}$ then any protocol for ϕ must produce a contextual privacy violation for some type profile in $\hat{\boldsymbol{\Theta}}$, that is, it must be that for any $P \in \mathcal{P}_{\mathfrak{S}}$, $\Gamma(P, \boldsymbol{\sigma}, \phi) \cap (\hat{\boldsymbol{\Theta}} \times N) \neq \emptyset$.

Proof. We first show that contextual privacy implies that no such set $\hat{\boldsymbol{\Theta}}$ exists. We prove this statement in the contrapositive. Hence, assume that $\hat{\boldsymbol{\Theta}}$ exists. As $\phi|_{\hat{\boldsymbol{\Theta}}}$ is non-constant (by hypothesis (i)) and P is a protocol for ϕ, the set of nodes that distinguish two type profiles in $\hat{\boldsymbol{\Theta}}$ is non-empty. Let v be any earliest (*i.e.* minimal in precedence order) node that distinguishes two type profiles in $\hat{\boldsymbol{\Theta}}$. The query at node v, s_v, must have $|s_v(\hat{\boldsymbol{\Theta}})| \geq 2$ as it distinguishes types in $\hat{\boldsymbol{\Theta}}$. Hence, by the hypothesis (ii) in the statement, there exist $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$ and $\boldsymbol{\theta}' = (\theta'_i, \boldsymbol{\theta}_{-i})$ with $s_v(\boldsymbol{\sigma}(\boldsymbol{\theta}, v)) \neq s_v(\boldsymbol{\sigma}(\boldsymbol{\theta}', v))$ and $\phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}')$. These constitute a contextual privacy violation.

For the converse direction, assume that no such $\hat{\boldsymbol{\Theta}}$ exists. We construct a contextually private protocol inductively. For the base case, we consider $\boldsymbol{\Theta}_r = \boldsymbol{\Theta}$. For the inductive step, we consider an arbitrary node v associated with $\boldsymbol{\Theta}_v = \hat{\boldsymbol{\Theta}}$. Either, $\phi|_{\hat{\boldsymbol{\Theta}}}$ is constant, and the protocol can terminate and compute ϕ, or not. If $\phi|_{\hat{\boldsymbol{\Theta}}}$ is not constant then there must be a query s'_v and a strategy profile $\boldsymbol{\sigma}$ such that $|s'_v(\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}, v))| \geq 2$ because, by assumption, there is an $\mathfrak{S}^*$-protocol for ϕ. Since hypothesis (i) holds, hypothesis (ii) cannot: there must be a query s_v such that $|s_v(\boldsymbol{\sigma}(\hat{\boldsymbol{\Theta}}, v))| \geq 2$ and there are no types $(\theta_i, \boldsymbol{\theta}_{-i}), (\theta'_i, \boldsymbol{\theta}_{-i}) \in \hat{\boldsymbol{\Theta}}$ with $s_v(\boldsymbol{\sigma}((\theta_i, \boldsymbol{\theta}_{-i}), v)) \neq s_v(\boldsymbol{\sigma}((\theta'_i, \boldsymbol{\theta}_{-i}), v))$ and $\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta'_i, \boldsymbol{\theta}_{-i})$. Hence, the query s_v does not introduce any contextual privacy violations. The induction terminates after finitely many rounds (because $|\boldsymbol{\Theta}| < \infty$) and the cardinality of $\boldsymbol{\Theta}_v$ strictly decreases along paths on the tree P. □

6.A.2 Proof of Proposition 6.2.

Proposition 6.2. *A choice function ϕ fails to be $\mathfrak{S}_{\mathrm{IE}}$-contextually private if and only if there exists some cylinder set $\hat{\boldsymbol{\Theta}} = \bigtimes_{i=1}^{n} \hat{\Theta}_i$ such that (i) $\phi|_{\hat{\boldsymbol{\Theta}}}$ is non-constant and (ii) for all agents i and all $\theta_i, \theta'_i \in \hat{\boldsymbol{\Theta}}_i$, θ_i and θ'_i are inseparable with respect to $\hat{\boldsymbol{\Theta}}$ and ϕ.*

In addition, it must be that for any protocol $(P, \boldsymbol{\sigma}) \in \mathcal{P}_{\mathfrak{S}_{\mathrm{IE}}}$, one of the agents i that has at least two inseparable types must incur a privacy violation, that is, $\Gamma(P, \boldsymbol{\sigma}, \phi) \cap \{(i, \boldsymbol{\theta}) \in N \times \boldsymbol{\Theta} \mid \exists \theta'_i \neq \theta_i : (\theta_i, \boldsymbol{\theta}_{-i}), (\theta'_i, \boldsymbol{\theta}_{-i}) \in \hat{\boldsymbol{\Theta}}\} \neq \emptyset$.

Proof. We first consider necessity. Assume for contradiction that there is a contextually private protocol P for the choice function ϕ and that there is a product set $\hat{\boldsymbol{\Theta}}$ such that all types are inseparable under $\hat{\boldsymbol{\Theta}}$ and ϕ is non-constant on this set.

As ϕ is non-constant on $\hat{\boldsymbol{\Theta}}$, the protocol must make a query separating type profiles $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta'_i, \boldsymbol{\theta}_{-i})$ for some agent i. Consider the earliest such query in the precedence order on P.

By the choice of v and $\theta_i \sim_{i,\phi,\hat{\boldsymbol{\Theta}}} \theta'_i$, there must be a chain $\theta_i^1, \theta_i^2, \ldots, \theta_i^k$ such that $\theta_i^1 = \theta_i$ and $\theta_i^k = \theta'_i$ and

$$\theta_i^1 \sim'_{i,\phi,\hat{\boldsymbol{\Theta}}} \theta_i^2 \sim'_{i,\phi,\hat{\boldsymbol{\Theta}}} \cdots \sim'_{i,\phi,\hat{\boldsymbol{\Theta}}} \theta_i^k.$$

That is, there is a chain of direct inseparability from θ_i to θ'_i. As θ_i and θ'_i are separated at v, there must be $l = 1, 2, \ldots, k-1$ such that v distinguishes θ_i^l and θ_i^{l+1}. By direct inseparability, there is $\boldsymbol{\theta}_{-i}$ such that $(\theta_i^l, \boldsymbol{\theta}_{-i}), (\theta_i^{l+1}, \boldsymbol{\theta}_{-i}) \in \hat{\boldsymbol{\Theta}}$ and $\phi(\theta_i^l, \boldsymbol{\theta}_{-i}) = \phi(\theta_i^{l+1}, \boldsymbol{\theta}_{-i})$. Together, these two observations yield a contradiction to contextual privacy of P.

Now consider sufficiency. We define a contextually private protocol inductively on nodes v. Throughout the induction, the following holds:

$$\text{For any terminal nodes } w, w' \text{ whose latest common ancestor in } P \text{ is } v \text{, there are no } (\theta_i, \boldsymbol{\theta}_{-i}) \in \boldsymbol{\Theta}_w \text{ and } (\theta'_i, \boldsymbol{\theta}_{-i}) \in \boldsymbol{\Theta}_{w'} \text{ such that } \phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta'_i, \boldsymbol{\theta}_{-i}). \tag{6.4}$$

Note that a protocol that satisfies (6.4) at all internal nodes is contextually private. We start the induction in the root node r of P.

Assume a protocol has been constructed until query v associated to type set $\boldsymbol{\Theta}_v \subseteq \boldsymbol{\Theta}$. If ϕ is constant on the remaining set, the node is terminal, and the outcome of the social choice function can be determined.

Otherwise, because there is no restriction to a product set $\hat{\boldsymbol{\Theta}}$ under which all types are inseparable, there are types θ_i, θ'_i that are separable under $\boldsymbol{\Theta}_v$ for agent i. Consider the binary query that distinguishes the equivalence class of θ_i, $[\theta_i]_{i,\phi,\boldsymbol{\Theta}_v}$, from its complement $\Theta_{v,i} \setminus [\theta_i]_{i,\phi,\boldsymbol{\Theta}_v}$, which is non-empty as it contains at least θ'_i. By definition of $\sim$, for any $\boldsymbol{\theta}_{-i}$, and any $\tilde{\theta}_i, \tilde{\theta}'_i$ such that $(\tilde{\theta}_i, \boldsymbol{\theta}_{-i}), (\tilde{\theta}'_i, \boldsymbol{\theta}_{-i}) \in \hat{\boldsymbol{\Theta}}$,

$$\phi(\tilde{\theta}_i, \boldsymbol{\theta}_{-i}) \neq \phi(\tilde{\theta}'_i, \boldsymbol{\theta}_{-i}),$$

completing the induction step and the proof (as (6.4) continues to hold). □

6.A.3 Proof of Theorem 6.1

Theorem 6.1. *Let ϕ be a social choice function, and consider a type profile $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. If for any subset $A \subseteq N$ of agents, and types $\theta_i' \in \Theta$ for agents $i \in A$,*

$$\phi(\boldsymbol{\theta}_A, \boldsymbol{\theta}_{-A}) \neq \phi(\boldsymbol{\theta}_A', \boldsymbol{\theta}_{-A}) \qquad \text{(collective pivotality)}$$

and for all $i \in A$

$$\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i}) \qquad \text{(no individual pivotality)}$$

then for any individual elicitation protocol $P \in \mathcal{P}_{\mathfrak{S}_{\mathrm{IE}}}$, there exists an agent $i \in A$ whose contextual privacy is violated at $\boldsymbol{\theta}$, i.e. $A \times \{\boldsymbol{\theta}\} \cap \Gamma(P, \phi) \neq \emptyset$.

Proof. Assume that $\phi(\boldsymbol{\theta}_A, \boldsymbol{\theta}_{-A}) \neq \phi(\boldsymbol{\theta}_A', \boldsymbol{\theta}_{-A})$ and $\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i})$ for all $i \in A$. Note that the cylinder set

$$\hat{\boldsymbol{\Theta}} := \bigtimes_{i \in A} \{\theta_i, \theta_i'\} \times \{\boldsymbol{\theta}_{-A}\}$$

satisfies the conditions of Proposition 6.2. Hence, we know that there is a contextual privacy violation for some type profile $\boldsymbol{\theta}' \in \hat{\boldsymbol{\Theta}}$.

To show that there is a violation at $\boldsymbol{\theta}$ in particular, consider a first query separating types from $\hat{\boldsymbol{\Theta}}$. This query must distinguish θ_i and θ_i' for some agent $i \in A$. By no individual pivotality, this violates contextual privacy for this agent at $\boldsymbol{\theta}$. □

6.A.4 Proof of Proposition 6.4.

Proposition 6.4. *Assume $|C| \geq 2$ and $|N| \geq 2$ and that there is no oversupply. Then any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\mathrm{IE}}}$ for a stable choice rule produces contextual privacy violations.*

Proof. The proof constructs collective but not individual pivotality.

Let $s_1 > s_2 > s_3 > s_4$. Fix the type profile of all $n - 2$ agents that are not i or j to be $\boldsymbol{\theta}_{-ij}$ where each agent has a score greater than s_1 for their top choice object, and their top choice object has capacity to accommodate them. Assume further that the remaining spots are for different objects. Label these objects with remaining spots a and b.

Consider the final two agents $i, j \in N$. Choose their type profiles to be:

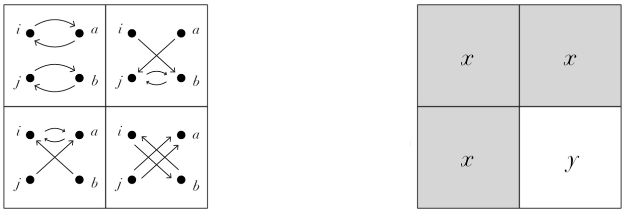
$$\begin{aligned}
\theta_i &= (a \succ_i b, s_i(a) = s_1, s_j(b) = s_4), & \theta_i' &= (b \succ_i a, s_i(a) = s_3, s_j(b) = s_2) \\
\theta_j &= (b \succ_j a, s_j(a) = s_4, s_j(b) = s_1), & \theta_j' &= (a \succ_j b, s_j(a) = s_2, s_j(b) = s_3).
\end{aligned}$$

The preferences of agents i and j for objects a and b are represented in Figure 6.6. Agent i and j's scores and preferences for other schools are arbitrary.

Let x be the outcome in which agent i is matched to school a and j is matched to b. Let y be the outcome in which agent i is matched to b and j is matched to a. In both x and y, all agents not i or j are assigned to their top choice object at which they have a high score.

Stability requires that

$$\phi(\theta_i, \theta_j, \boldsymbol{\theta}_{-ij}) = (\theta_i, \theta_j', \boldsymbol{\theta}_{-ij}) = (\theta_i', \theta_j, \boldsymbol{\theta}_{-ij}) = x$$



x	x
x	y

Figure 6.6: Constructing collective but not individual pivotality for college assignment. Agent types $\theta_i, \theta_i', \theta_j, \theta_j'$ (left, arrows from agents denote favored object, arrows from objects denote high score); outcomes under any stable choice rule (right, where $x = ((i,a),(j,b))$ and $y = ((j,a),(i,b))$).

and

$$\phi(\theta_i, \theta_j, \boldsymbol{\theta}_{-ij}) = y.$$

We have chosen a particular $\boldsymbol{\theta}_{-ij} \in \boldsymbol{\Theta}_{-ij}$, and particular $\theta_i, \theta_i', \theta_j, \theta_j' \in \Theta$ such that the condition of Theorem 6.1 holds. □

6.A.5 Strengthening of Proposition 6.5 Under No Ties.

Proposition 6.2 allows us to strengthen Proposition 6.5 for cases without ties in the type profile (*i.e.* where no two agents have the same type).

Proposition 6.11. *Assume $|\Theta| \geq 9$. If there are $n \geq 3$ agents, any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\mathrm{IE}}}$ for the second-price auction rule produces contextual privacy violations.*

Proof. Consider the three highest priority (in lexicographic order) agents 1, 2 and 3, and 9 types

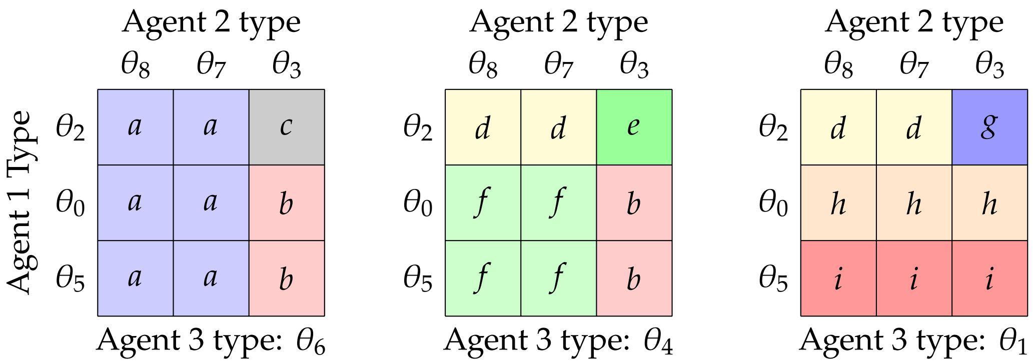
$$\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8,$$

that are increasing in the index.

Consider the product set $\hat{\boldsymbol{\Theta}} = \{\theta_5, \theta_0, \theta_2\} \times \{\theta_8, \theta_7, \theta_3\} \times \{\theta_6, \theta_4, \theta_1\} \times \bigtimes_{i=4}^{n}\{\theta_0\}$. In this product set, the first factor represents types of agent 1, the second represents possible types of agent 2, the third represents possible types of agent 3, and all other agents have type θ_0. We will show that when ϕ^{SPA} is evaluated on this restricted product set, it is non-constant and all types in the product set are inseparable.

To see this, we construct the tensor of outcomes for the product set. This tensor is represented in Figure 6.7. We represent agent 1's type on the up-down axis, agent 2's type on the left-right axis, and agent 3's type is constant for each box. The outcomes under ϕ^{SPA} are represented by letters and colors. For example, the upper left corner in the left-most box signifies $\phi^{\mathrm{SPA}}(\theta_2, \theta_8, \theta_6) = a$, where a is the outcome under which agent 2 wins the object and pays a price θ_6.

To see that this constitutes a violation of contextual privacy, we show that: (i) ϕ is non-constant on $\hat{\boldsymbol{\Theta}}$, and (ii) for all agents i, and all $\theta_i, \theta_i' \in \hat{\boldsymbol{\Theta}}$, θ_i and θ_i' are inseparable. As for (i), we can observe immediately that $\phi|_{\hat{\boldsymbol{\Theta}}}$ is non-constant. To see (ii) that all types are inseparable, we go through each agent in turn.



Agent 3 type: θ_6

Agent 1 Type \ Agent 2 type	θ_8	θ_7	θ_3
θ_2	a	a	c
θ_0	a	a	b
θ_5	a	a	b

Agent 3 type: θ_4

Agent 1 Type \ Agent 2 type	θ_8	θ_7	θ_3
θ_2	d	d	e
θ_0	f	f	b
θ_5	f	f	b

Agent 3 type: θ_1

Agent 1 Type \ Agent 2 type	θ_8	θ_7	θ_3
θ_2	d	d	g
θ_0	h	h	h
θ_5	i	i	i

Figure 6.7: Counterexample ϕ^{SPA} with $n = 3$ and $|\Theta| = 9$.

- Agent 1: Outcome a is the same for $\boldsymbol{\theta}_{-1} = (\theta_7, \theta_6)$, hence all agent 1 types are inseparable.
- Agent 2: Outcome i for $\boldsymbol{\theta}_{-2} = (\theta_5, \theta_1)$ show that all agent 2 types are inseparable.
- Agent 3: θ_6 and θ_4 are inseparable because they both yield outcome b for $\boldsymbol{\theta}_{-3} = (\theta_0, \theta_3)$. θ_1 and θ_4 are inseparable because they both yield outcome d for $\boldsymbol{\theta}_{-3} = (\theta_2, \theta_7)$.

□

6.A.6 Proof of Proposition 6.6.

Proposition 6.6. *Assume $|\Theta| \geq 4$ and $|N| \geq 4$. Any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$ for an efficient, uniform-price double auction price rule produces contextual privacy violations for $\boldsymbol{\theta}$ where two agents determine the clearing price.*

Proof. We construct collective but not individual pivotality. Consider four agents i, j, k, and ℓ, two types $\underline{\theta}, \bar{\theta} \in \Theta$ and a partial type profile $\boldsymbol{\theta}_{-ijk\ell}$ such that

$$\text{median}(\boldsymbol{\theta}_{-ijk\ell}) = \{\underline{\theta}, \bar{\theta}\}.$$

Endowments are arbitrary.

Consider outcomes in type profiles $(\theta_i, \theta_j, \theta_k, \theta_\ell, \boldsymbol{\theta}_{-ijk\ell})$ as represented on the left side in Figure 6.8.

Denote $\underline{\theta}$ the outcome in which the market clearing price is $\underline{\theta}$ and let $\overline{\theta}$ be the outcome in which the market clearing price is $\overline{\theta}$. Consider first the top square which holds the types of agents i and ℓ fixed and varies the types of agents j and k. Efficiency requires $\phi(\underline{\theta}, \underline{\theta}, \underline{\theta}, \underline{\theta}, \boldsymbol{\theta}_{-ijk\ell}) = \phi(\underline{\theta}, \underline{\theta}, \overline{\theta}, \underline{\theta}, \boldsymbol{\theta}_{-ijk\ell}) = \phi(\underline{\theta}, \overline{\theta}, \underline{\theta}, \underline{\theta}, \boldsymbol{\theta}_{-ijk\ell}) = \underline{\theta}$.

Now consider the bottom square which holds the types of agents j and k fixed and varies the types of agents i and ℓ. Efficiency requires that $\phi(\bar{\theta}, \bar{\theta}, \bar{\theta}, \underline{\theta}, \boldsymbol{\theta}_{-ijk\ell}) = \phi(\underline{\theta}, \bar{\theta}, \bar{\theta}, \bar{\theta}, \boldsymbol{\theta}_{-ijk\ell}) = \phi(\bar{\theta}, \bar{\theta}, \bar{\theta}, \bar{\theta}, \boldsymbol{\theta}_{-ijk\ell}) = \overline{\theta}$.

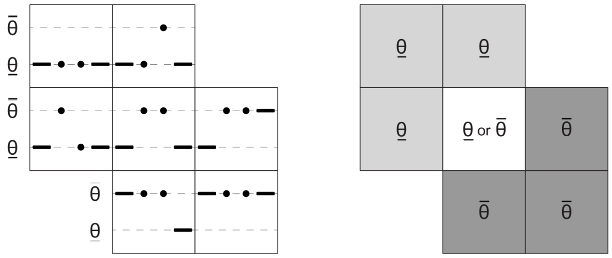


Figure 6.8: Constructing collective but not individual pivotality for the double auction. Combinations of types for agents i, j, k, ℓ (left); outcomes in an efficient uniform-price choice rule (right).

The outcome under the type profile in the box that conjoins the two squares,

$$(\underline{\theta}, \bar{\theta}, \bar{\theta}, \underline{\theta}, \boldsymbol{\theta}_{-ijk\ell})$$

must be *either* $\underline{\theta}$ or $\overline{\theta}$. If it is $\underline{\theta}$, then we can apply Theorem 6.1 for the bottom k-ℓ-square. If it is $\overline{\theta}$, then we can apply this statement in the top i-j-square. □

6.A.7 Proof of Proposition 6.7.

Proposition 6.7. *Assume $|N| \geq 2$ and $|X| \geq 2$ and the preference domain of single-peaked preferences. Any protocol $P \in \mathcal{P}_{\mathfrak{S}_{\mathrm{IE}}}$ for a generalized median voting rule that is neither the maximum nor the minimum rule produces contextual privacy violations.*

Proof. The proof is related to the one for Proposition 6.6, however simplified as the median is unique for an odd number of inputs, as is the case for a generalized median voting rules. Consider two adjacent types $\theta, \theta' \in \Theta$ and $\boldsymbol{\theta}_{-ij} \in \boldsymbol{\Theta}_{-ij}$ such that

$$\theta = \text{median}(\boldsymbol{\theta}_{-ij}, k_1, k_2, \ldots, k_{n-1})$$

and exactly one type in $\boldsymbol{\theta}_{-ij}, k_1, k_2, \ldots, k_{n-1}$ is θ. The type profiles $(\theta, \theta, \boldsymbol{\theta}_{-ij})$, $(\theta', \theta, \boldsymbol{\theta}_{-ij})$ and $(\theta, \theta', \boldsymbol{\theta}_{-ij})$ will all result in outcome θ. However, $(\theta', \theta', \boldsymbol{\theta}_{-ij})$ will result in θ'. This hence produces an instance of collective pivotality without individual pivotality and shows that there must be a contextual privacy violation for at least one of the agents i, j. □

6.A.8 Proof of Theorem 6.2.

Theorem 6.2. *Let $(P, \boldsymbol{\sigma})$ be a $\mathfrak{S}_{\mathrm{IE}}$-protocol for ϕ, where ϕ exhibits interval pivotality. There is a bimonotonic protocol $(P', \boldsymbol{\sigma}')$ that has the same set of contextual privacy violations as $(P, \boldsymbol{\sigma})$,* i.e. $\Gamma(P, \boldsymbol{\sigma}, \phi) = \Gamma(P', \boldsymbol{\sigma}', \phi)$.

Proof. We begin with an arbitrary protocol and transform it with two operations into a bimonotonic protocol that has the same set of contextual privacy violations: *filling-in* and *scrubbing*. By the definition of individual elicitation, we can identify queries with partitions of an agent's type space.

We begin with an auxiliary definition that will be used in the anchoring and filling-in steps.

Definition 6.9 (v-before-v' injected protocol)**.** Let $P = (V, E)$ be a protocol. Let $s_v\colon \mathbf{M} \to \text{children}(v)$ be a query and $\sigma_{iv}\colon \Theta \to M$ be a strategy for the queried agent. Let $v' \in V$ be a non-root node with parent u. Denote subtree(v') to be the subtree from v' in P. Define the v-before-v'-injected protocol $P_{v,v'}$ to be the protocol in which u has a single child v (children(u) = $\{v\}$) and all children of v are followed by subtree(v').

Note that subtree(v') refers to the sequence of queries s_w for nodes w in the subtree following v' in the original protocol P. When we inject v before v' in the protocol $P_{v,v'}$, all children of v are followed by the sequence of queries s_w from the original protocol P. See Figure 6.9 for an illustration of injection.

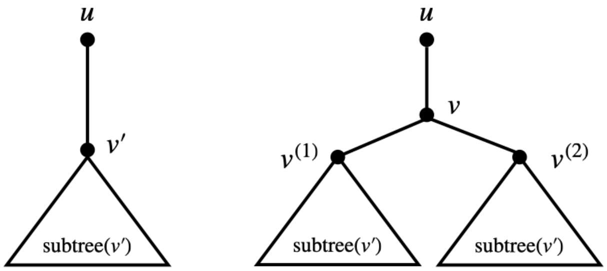


Figure 6.9: Protocol P (left); v-before-v' injected protocol $P_{v',v}$ (right). Here, in $P_{v,v'}$, the nodes $\{v^{(1)}, v^{(2)}\}$ = children(v) are "clones" of v' in original protocol P.

We denote $i(v)$ as the agent who receives a query at node v.

Definition 6.10. Let Θ be a finite ordered type space and denote by succ(θ) the adjacent type higher than θ in Θ (if it exists) and by pred(θ) the adjacent type lower than θ in Θ (if it exists). Also, let s_v be a query for node v. We call $\underline{\theta}_{v,i}$ the *lowest separator* if

$$\underline{\theta}_{v,i} = \min\{\theta_i \in \Theta \mid \exists \boldsymbol{\theta}_{-i} : (\theta_i, \boldsymbol{\theta}_{-i}) \in \boldsymbol{\Theta}_v \text{ and } s_v(\sigma_i(\text{succ}(\theta_i), v), \boldsymbol{\sigma}_{-i}(\boldsymbol{\theta}_{-i}, v)) \neq s_v(\sigma_i(\theta_i, v), \boldsymbol{\sigma}_{-i}(\boldsymbol{\theta}_{-i}, v))\}$$

and we call $\overline{\theta}_{v,i}$ the *highest separator* if

$$\overline{\theta}_{v,i} = \max\{\theta_i \in \Theta \mid \exists \boldsymbol{\theta}_{-i} : (\theta_i, \boldsymbol{\theta}_{-i}) \in \boldsymbol{\Theta}_v \text{ and } s_v(\sigma_i(\text{pred}(\theta_i), v), \boldsymbol{\sigma}_{-i}(\boldsymbol{\theta}_{-i}, v)) \neq s_v(\sigma_i(\theta_i, v), \boldsymbol{\sigma}_{-i}(\boldsymbol{\theta}_{-i}, v))\}.$$

That is, the lowest separator at node v is the lowest element in the type space that is distinguished from its next lowest type at v. The highest separator at node v is the highest element in the type space that is distinguished from its next highest type at node v. If these sets are empty, the query is trivial in that it does not affect the knowledge of the designer. It is then without loss to drop this query from the protocol.

Lemma 6.2. *Let ϕ be interval pivotal, $i(v) = i(v')$, and $v' \in \operatorname{subtree}(v)$. Define*

$$\begin{aligned} \underline{\theta}_{v,v'} &= \min\{\underline{\theta}_{v,i}, \underline{\theta}_{v',i}\} \\ \overline{\theta}_{v,v'} &= \max\{\overline{\theta}_{v,i}, \overline{\theta}_{v',i}\} \end{aligned} \tag{6.5}$$

where $\underline{\theta}_{v,i}, \overline{\theta}_{v,i}$ are the highest and lowest separators at v as defined above (and similarly for v'). Then, for any $\tilde{\theta} \in [\underline{\theta}_{v,v'}, \overline{\theta}_{v,v'}]$, a threshold query *$v''$ with threshold $\tilde{\theta}$ satisfies*

$$\Gamma(P, \boldsymbol{\sigma}, \phi) = \Gamma(P_{v'',v'}, \boldsymbol{\sigma}', \phi),$$

where $\boldsymbol{\sigma}'$ is a threshold strategy with threshold $\tilde{\theta}$ at v'' and the same as $\boldsymbol{\sigma}$ on all other nodes and all other agents.

In other words, consider an agent that is asked two queries v and v' one after the other on one path through the protocol. Inserting a threshold query *in between* v and v', where the threshold $\tilde{\theta}$ lies *between* the highest and lowest types that v and v' can distinguish, this insertion will result in no change to the set of contextual privacy violations.

Proof. We show first that $\Gamma(P_{v'',v'}, \boldsymbol{\sigma}', \phi) \subseteq \Gamma(P, \boldsymbol{\sigma}, \phi)$. Let $(i, \boldsymbol{\theta}) \in \Gamma(P_{v'',v'}, \boldsymbol{\sigma}', \phi) \backslash \Gamma(P, \boldsymbol{\sigma}, \phi)$ for contradiction. Then there is a θ_i' such that $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$ are distinguished at v'' and

$$\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i}).$$

By symmetry, it is without loss to assume $\theta_i \leq \theta_i'$. As θ_i and θ_i' are distinguished, and by the definition of threshold queries, it must be that $\theta_i \leq \tilde{\theta} < \theta_i'$.

We will now show that $\boldsymbol{\theta}$ has a contextual privacy violation for agent i at node v or node v' in P, leading to a contradiction. As $\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i})$ and because of interval pivotality, there are two cases, corresponding to the "upper" or "lower" interval over which $\theta_i \mapsto \phi(\theta_i, \boldsymbol{\theta}_{-i})$ is constant:

(a) $\phi(\hat{\theta}_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i})$ for all types $\hat{\theta}_i \leq \theta_i'$ (in particular this holds for $\hat{\theta}_i = \underline{\theta}_{v,v'}$ by the definition of the lowest separator), or

(b) $\phi(\hat{\theta}_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i, \boldsymbol{\theta}_{-i})$ for all types $\hat{\theta}_i > \theta_i$ (in particular this holds for $\hat{\theta}_i = \overline{\theta}_{v,v'}$ by the definition of the highest separator).

For the first case (a), for notational simplicity call v the node that attains the minimum in (6.5). By definition of $\underline{\theta}_{v,v'}$, the types $\underline{\theta}_v, \operatorname{succ}(\underline{\theta}_v)$ are distinguished at v. There are two further cases within case (a):

- $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$ are distinguished at v. In this case, both type profiles produce contextual privacy violations for agent i at v.

- $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$ are *not* distinguished at v, but this means that θ_i is distinguished from either of $\underline{\theta}_v, \operatorname{succ}(\underline{\theta}_v)$, or both. So, $(i, \boldsymbol{\theta})$ produces a contextual privacy violation with either $(\underline{\theta}_v, \boldsymbol{\theta}_{-i})$ or $(\operatorname{succ}(\underline{\theta}_v), \boldsymbol{\theta}_{-i})$.

For the second case (b), we can follow similar reasoning, but with flipped inequality signs. In this case, one proceeds by showing that a contextual privacy violation is produced at the node that attains the maximum in (6.5).

For the converse direction, observe that $P_{v'',v'}$ reveals weakly more information to the designer than P. Hence, also $\Gamma(P, \boldsymbol{\sigma}, \phi) \subseteq \Gamma(P_{v'',v'}, \boldsymbol{\sigma}', \phi)$. □

Using Lemma 6.2, we will "fill-in" all queries between the highest and lowest separator. This results in a *filled-in* protocol.

Definition 6.11. We call a protocol P *filled-in* if (a) all queries are threshold queries and (b) all thresholds of adjacent queries (with respect to protocol P) are adjacent in Θ.

That is, in a filled-in protocol, every query is a threshold query and the threshold for every query adjacent in the protocol is also adjacent in the type space.

Lemma 6.3. *Let ϕ be an interval pivotal choice rule. Then, for any protocol $(P, \boldsymbol{\sigma})$, there is a filled-in protocol $(P', \boldsymbol{\sigma})$ such that $\Gamma(P, \boldsymbol{\sigma}, \phi) = \Gamma(P', \boldsymbol{\sigma}', \phi)$.*

Proof. We prove this lemma in three steps: anchoring, inserting and deleting.

Step 0: Grounding. We first add trivial queries $s_i(\boldsymbol{\theta}) = r_i$ to all agents in the beginning, where r_i can be thought of as a copy of the root node. These allow us to perform protocol injection on initial queries to agents and resolves issues for protocols where an agent only gets a single query. Inserting trivial queries affects neither computability of the choice rule nor contextual privacy violations.

Step 1: Anchoring. We then introduce threshold queries at the highest ($\overline{\theta}_{v,v'}$) and lowest ($\underline{\theta}_{v,v'}$) separators of adjacent queries v, v' to the same agent, that is, before the later in P of v, v', $i(v) = i(v')$. This leads to the introduction of at most $2|V|$ many new queries and does not affect computability of the choice rule or contextual privacy violations.

Step 2: Inserting. For any pair of threshold queries (with respect to the protocol P), v, v' to the same agent, whose thresholds are not adjacent in θ, insert threshold queries for all thresholds between thresh(v) and thresh(v') (in the type space) before the later of v, v'. This leads to at most $2|V||^2\Theta|$ many queries.

Step 3: Deleting. For any non-threshold queries v, all thresholds between $\underline{\theta}_v$ and $\overline{\theta}_v$ were added in Steps 1 and 2. This implies that all non-threshold queries can be deleted without affecting computability of the choice rule. Additionally, their deletion cannot affect contextual privacy violations because any distinguished types under the queries are distinguished at at least one of the added threshold queries.

The resulting protocol is filled-in. □

A filled-in protocol is almost bimonotonic. But, it may have sequences of threshold queries to a single agent that increase and then decrease (or vice versa) in the thresholds. So, we use a *scrubbing* operation to dispose of non-monotonicity in the threshold queries.

Let $(P, \boldsymbol{\sigma})$ be a protocol. A query v that distinguishes no type profiles in $\boldsymbol{\Theta}_v$ is *trivial*. (An example of trivial queries are duplicates, queries for which there is an earlier, identical query).

To create a protocol without trivial queries, we delete such queries "from the bottom up", *i.e.* from the leaves to the root. We call this operation *scrubbing*. For any trivial query v, there is a unique $w \in \text{children}(v)$ such that $\boldsymbol{\Theta}_w \neq \emptyset$. To delete v, attach subtree(w) to parent(v) and delete subtree(w') for all $w' \in \text{children}(v) \setminus \{w\}$. This operation terminates after at most $|V(P)|$ rounds, and neither affects computability of the choice rule nor contextual privacy violations. The scrubbing operation is illustrated in Figure 6.10.

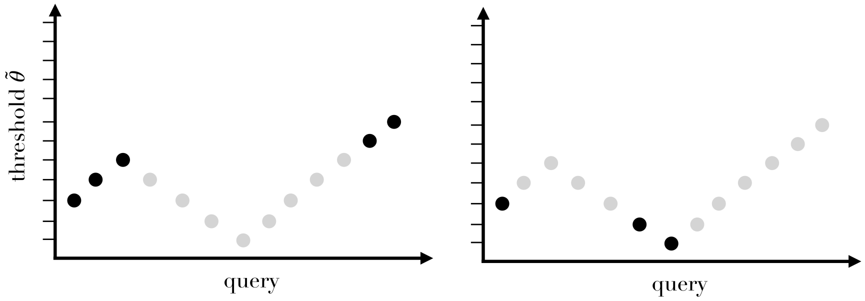


Figure 6.10: Queries to agent i in protocol P, scrubbed after affirmative answer to initial query (left); Queries to agent i in protocol P scrubbed after negative answer to initial query (right). The x-axis represents the order of the queries in protocol P. The y-axis is the threshold of threshold queries in protocol P. Each query in the original protocol is a dot, queries deleted during scrubbing are gray dots.

This results in a bimonotonic protocol. An agent's response to the initial query tells us that either all queries with higher thresholds are trivial, or that all queries with lower thresholds are trivial. Consider the former case, the latter is analogous. The agent's initial response indicates that they are below the initial threshold. Because the protocol is filled-in, the next non-trivial query must have a threshold below and adjacent to the initial threshold. The response to this query either fully reveals the agent's type (in which case all future queries are trivial) or renders any threshold above this one trivial. This continues until the protocol terminates, leading to a sequence of queries with monotonically decreasing thresholds.

□

6.A.9 Proof of Theorem 6.3.

Theorem 6.3. *For the second-price auction rule ϕ^{SPA}, there is:*

(a) a unique minimal set of contextual privacy violations for W- *and* LPL*-protecting $\mathfrak{S}_{\text{IE}}$-protocols,*

(b) a unique minimal set of contextual privacy violations for L- *and* HPW*-protecting $\mathfrak{S}_{\text{IE}}$-protocols.*

Two bimonotonic protocols attain these sets of contextual privacy violations, which we call the ascending-join protocol *and the* overdescending-join protocol*, respectively. Both of these protocols are maximally contextually private among $\mathfrak{S}_{\text{IE}}$-protocols for ϕ^{SPA}.*

We prove a more general version of Theorem 6.3 that clarifies the structure of the argument, and shows that our proof technique applies more broadly. In particular, we show that for any k-item uniform price auction where the price is set by the $(k+1)^{\text{st}}$ agent's value, there is a unique maximal protocol for contextual privacy.

We will denote k-item uniform $(k+1)^{\text{st}}$ price auction rules by $\phi^{k\text{-PA}}$. Before stating our more general result, we must define protection sets for $\phi^{k\text{-PA}}$ that are analogous to the protection sets defined for ϕ^{SPA} from the main text. We call an agent a *winner* if she is allocated a good. If she also does not determine the price, then she is a *loser*. For the $(k+1)^{\text{st}}$-price auction, we define the sets of winners and losers as follows:

$$\begin{aligned} \mathrm{W} &= \{(i, \boldsymbol{\theta}) \mid i \text{ is a winner at } \boldsymbol{\theta}\} \\ \mathrm{L} &= \{(i, \boldsymbol{\theta}) \mid i \text{ is neither a winner nor determines the price at } \boldsymbol{\theta}\}. \end{aligned}$$

We also have subsets of the winner and loser protection sets, which are exactly the same as the main text—we reproduce the definitions here for convenience. Let $d(\boldsymbol{\theta})$ be the index of the agent who determines the price at type profile $\boldsymbol{\theta}$. We define the protection sets LPL and HPW as follows:

$$\begin{aligned} \mathrm{LPL} &= \mathrm{L} \cap \{(i, \boldsymbol{\theta}) \mid i > d(\boldsymbol{\theta})\} \\ \mathrm{HPW} &= \mathrm{W} \cap \{(i, \boldsymbol{\theta}) \mid i < d(\boldsymbol{\theta})\}. \end{aligned}$$

The generalization of Theorem 6.3 is as follows.

Theorem 6.4. *For the uniform $(k+1)^{st}$-price k-item auction rule $\phi^{k\text{-PA}}$, there is:*

(a) a unique minimal set of contextual privacy violations for W- *and* LPL-*protecting $\mathfrak{S}_{\text{IE}}$-protocols,*
(b) a unique minimal set of contextual privacy violations for L- *and* HPW-*protecting $\mathfrak{S}_{\text{IE}}$-protocols.*

Two bimonotonic protocols attain these sets of contextual privacy violations, which we call the ascending-join protocol *and the* overdescending-join protocol, *respectively. Both of these protocols are maximally contextually private among $\mathfrak{S}_{\text{IE}}$-protocols for $\phi^{k\text{-PA}}$.*

Proof. The proofs for both (a) and (b) are constructive. We show that at each point in the protocol, the principal has a unique choice for which agent to ask next, and for which threshold. This relies on our characterization theorem Theorem 6.2.

Our argument will roughly structure the protocol into "rounds," after which the designer's information has a particular structure. Whenever such a structure is reached, a new round begins.

We begin with part (a). The relevant designer information structure for the ascending-join protocol: W- *and* LPL-*protective information states.*

Definition 6.12. We call a set of type profiles $\boldsymbol{\Theta}_v \subseteq \boldsymbol{\Theta}$ a W- and LPL-*protective information state* if there is a type $\tilde{\theta} \in \Theta$ such that (recall that $\Theta_{v,i}$ is the projection of $\boldsymbol{\Theta}_v$ onto the i^{th} component):

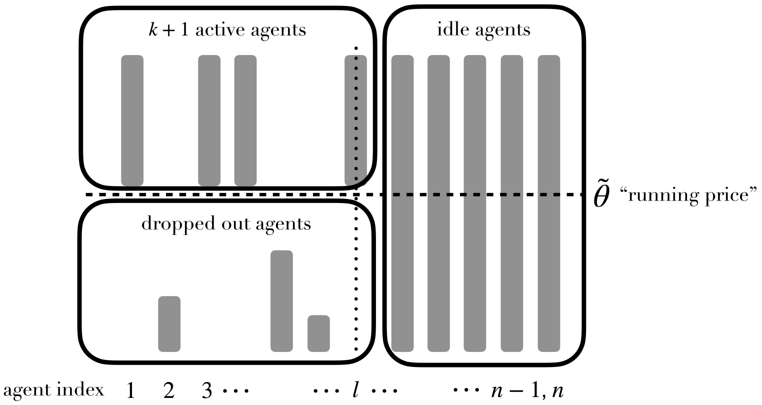

$k+1$ active agents
idle agents
$\tilde{\theta}$ "running price"
dropped out agents
agent index 1 2 3 ⋯ ⋯ l ⋯ ⋯ $n-1, n$

Figure 6.11: W- and LPL-Protective Information State

1. There are exactly $k+1$ agents i such that $\Theta_{v,i} = \{\theta_i \mid \theta_i > \tilde{\theta}\}$.
2. For agents $i = 1, 2, \ldots, l$ with l being the highest index among agents who ever got a question (*i.e.* $l := \max\{l \colon \Theta_{v,l} \neq \Theta\}$), it must be that, either:
   (i) $\Theta_{v,i} = \{\theta_i \mid \theta_i > \tilde{\theta}\}$ (*"dropped-out agent"*), or
   (ii) $\Theta_{v,i} \subseteq \{\theta_i \mid \theta_i \leq \tilde{\theta}\}$ (*"active agent"*).

We call all other agents *idle*.

For a protective information state $\boldsymbol{\Theta}_v$, we call $\tilde{\theta}$ the *running price*. Sets that are protective have $k+1$ agents about whom it is known only that their type is strictly above the running price (active agents). Nothing is known about agents with indices "to the right" of the agent with the highest index among agents who have ever been asked a question (idle agents). All other agents have a type weakly below the running price (dropped-out agents). We illustrate this definition in Figure 6.11.

Definition 6.13. We say that a query v is ϕ-redundant if for any θ_i, θ'_i, $\boldsymbol{\theta}_{-i}$ such that $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta'_i, \boldsymbol{\theta}_{-i})$ are distinguished at v, we have

$$\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta'_i, \boldsymbol{\theta}_{-i}).$$

We make the following observation, which is direct from the definitions.

Lemma 6.4. *Let P be a protocol for ϕ and v be a ϕ-redundant query in P. Any protocol P' that removes v and continues with* any *of the children of v from P is still a protocol for ϕ. P' is weakly more contextually private than P.*

We next make an additional observation about the structure of protection sets to more easily reason about which queries are forbidden under a given protection set. To this end, the structure of a *closed* protection set is helpful.

Definition 6.14 (Closed protection sets)**.** We call a protection set A *downward closed* if we have that for $(i, (\theta_i, \boldsymbol{\theta}_{-i})) \in A$ and $\theta'_i \leq \theta_i$,

$$(i, (\theta'_i, \boldsymbol{\theta}_{-i})) \in A.$$

A protection set is A *upward closed* if we have that for $(i, (\theta_i, \boldsymbol{\theta}_{-i})) \in A$ and $\theta_i' \geq \theta_i$,

$$(i, (\theta_i', \boldsymbol{\theta}_{-i})) \in A.$$

If a model is downward closed or upward closed, we call it *closed*.

Definition 6.15 (ϕ-closed protection sets)**.** If A is a closed protection set, and for any $i \in N$, $\theta_i, \theta_i' \in \Theta$ and $\boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}_{-i}$, such that $(i, (\theta_i, \boldsymbol{\theta}_{-i})), (i, (\theta_i', \boldsymbol{\theta}_{-i})) \in A$, it holds

$$\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i}),$$

we call it ϕ-closed.

The property ϕ-closedness may be seen as a minimality property for $\mathfrak{S}_{\text{IE}}$ protocols: Any distinction of type profiles in a ϕ-closed protection set would lead to a contextual privacy violation for the queried agent.

Lemma 6.5. *Let A be a ϕ-closed protection set. A protocol P has a contextual privacy violation in A if it contains a query v and two type profiles $(\theta_i, \boldsymbol{\theta}_{-i}), (\theta_i', \boldsymbol{\theta}_{-i})$ that P distinguishes such that $(i, (\theta_i, \boldsymbol{\theta}_{-i})), (i, (\theta_i', \boldsymbol{\theta}_{-i})) \in A$.*

Proof. We show this lemma for upward closedness. The proof for downward closedness is analogous. We say that "i is in A under $\boldsymbol{\theta}$" if $(i, \boldsymbol{\theta}) \in A$.

Consider a query v in protocol P and two type profiles $\boldsymbol{\theta}, \boldsymbol{\theta}'$ that are distinguished at v such that i is in A for both $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$. Without loss we may assume that $\theta_i < \theta_i'$. As A is upward closed, it must be that i is in A also under $(\theta_i', \boldsymbol{\theta}_{-i})$, and does not change the outcome, by ϕ-closedness. This means that there is a contextual privacy violation in the protection set A. □

Note that the protection sets we defined in this section (that is, W, L, HPW, LPL) are $\phi^{k\text{-PA}}$-closed.

Lemma 6.6. *There is a unique sequence of non-redundant,* W- *and* LPL-*protecting queries on* $\boldsymbol{\Theta}$ *that leads to a* W- *and* LPL-*protective information state or termination.*

Proof. We may restrict to bimonotonic protocols by Theorem 6.2, in particular to protocols that only use threshold queries. By Lemma 6.5, any query with a threshold higher than the lowest type violates winner privacy because both a negative answer and an affirmative answer may lead to the queried agent winning. This continues to hold until k higher-priority agents have given an affirmative answer to a threshold query about the lowest type. We will check that the $(k+1)^{\text{st}}$ query will still have the lowest type as a threshold, and leads to a W- and LPL-protective information state. For the first k queries, we only need to determine the order in which agents are asked.

Denote $h \in \{1, 2, \ldots, n\}$ the number affirmative answers given so far. Call agents for which a type higher than the running price is not ruled out "remaining". We show by induction over h the following statement:

> The only non-redundant query that does not violate winner or low-priority loser privacy is to the $(k+1-h)^{\text{th}}$ remaining agent in priority order. (6.6)

Note that all queries to agents who dropped out are redundant and it is without loss to abstract from them by Lemma 6.4. Also note that all queries to agents $1, 2, \ldots, k - h$ make it possible that the queried agent is a winner both in the case of an affirmative and a negative answer, hence violating W privacy by Lemma 6.5. All queries to agents $k - h + 2, k - h + 3, \ldots, n$ make it possible that the queried agent is a low-priority loser in the case of an affirmative or a negative answer, hence violating LPL privacy. Asking the $(k + 1 - h)^{\text{th}}$ remaining agent does not lead to a W or LPL contextual privacy violation.

(As a concrete example, consider the very first query. We may not ask agent $1, 2, \ldots, k$ as they might win both with an affirmative and a negative answer. We may also not ask agents $k + 2, k + 3, \ldots, n$, as they might be low-priority losers in case of a positive and a negative answer. Hence, agent $k + 1$ must be queried. What's important is that a negative answer from the $(k+1)^{\text{st}}$ unambiguously means that they are not a winner. An affirmative answer from this agent means that they are not a low-priority loser.)

This determines the order of the first k queries. Observe that agents $2, \ldots, m$ for some $m \leq n$ have been asked by the property (6.6). If we show that agent 1 is the only agent that can be queried, and must be queried for the lowest type, then the resulting information of the designer is a W- and LPL-protective information state (unless the protocol can terminate before this is the case).

We distinguish three cases. First, asking agents $1, \ldots, m$ for a higher threshold might lead to them being a winner for both an affirmative and a negative answer rendering these queries forbidden by Lemma 6.5. Second, asking agents $m + 1, \ldots, n$ for any threshold might lead them to being a low-priority loser for both an affirmative or a negative answer, rendering these forbidden by Lemma 6.5. Third, asking agent 1 for any threshold higher than the lowest type might lead to them being a winner for both an affirmative and a negative answer, making this query forbidden, also by Lemma 6.5. Asking agent 1 a threshold query with the lowest type does not lead to W and LPL violations. Note that asking agent 1 complies with the agent order in (6.6). Hence, (6.6) can be seen as the definition of the ascending-join protocol.

Either the designer reaches information state $\mathbf{\Theta}_v$ in which $k + 1$ agents have given an affirmative answer. In this case, the designer's information state $\mathbf{\Theta}_v$ is W- and LPL-protective. Or, all agents except for k have given negative answers. In this case, the protocol may terminate (allocating to the remaining k agents at the current running price). □

Lemma 6.7. *On a* W- *and* LPL-*protective set* $\mathbf{\Theta}_v$, *there is a unique sequence of non-redundant,* W- *and* LPL-*protecting queries, that leads to a* W- *and* LPL-*protective set or termination.*

Proof. We reduce this to Lemma 6.6. Observe that any query to a dropped-out agent is redundant. We may therefore drop them from consideration. Also observe that any query with a threshold lower than the running price is either trivial (if it is to an active agent), redundant (if it is to a dropped-out agent) or could lead to the queried agent being a high-priority winner for both an affirmative and a negative answer (if it is to an idle agent). Hence, we may reduce to a situation in which the designer faces only active agents and needs to compute the choice rule $\phi^{k\text{-PA}}$ on a restricted type space $\Theta' := \{\theta' \mid \theta' > \tilde{\theta}\}$, where $\tilde{\theta}$ is the current running price, and apply Lemma 6.6. □

This concludes the proof of part (a). We now turn to part (b), using a similar proof strategy.

We first need to define an analogous version of a protective information state.

Definition 6.16. We call a set of type profiles $\mathbf{\Theta}_v \subseteq \mathbf{\Theta}$ a k-item L- and HPW-*protective information state* if there is a type $\tilde{\theta} \in \Theta$ such that (recall that $\Theta_{v,i}$ is the projection of $\mathbf{\Theta}_v$ onto the i^{th} component):

1. There are exactly $n - k + 1$ agents i such that $\Theta_{v,i} = \{\theta_i \mid \theta_i \leq \tilde{\theta}\}$.

2. For agents $i = l, \ldots, n$ with l being the smallest index among agents who ever got a question (*i.e.* $l \coloneqq \min\{l \colon \Theta_{v,l} \neq \Theta\}$), it must be that, either:

   (i) $\Theta_{v,i} \subseteq \{\theta_i \mid \theta_i > \tilde{\theta}\}$ (*"winners"*), or

   (ii) $\Theta_{v,i} = \{\theta_i \mid \theta_i \leq \tilde{\theta}\}$ (*"active agent"*).

We call all other agents *idle*.

Because of Theorem 6.2, we may restrict to bimonotonic protocols. As in part (a), we will show that there is a unique sequence of non-redundant queries within each round. The following two lemmas conclude this part of the proof. The first lemma shows that after a "round," the designer's information state is L- and HPW-protective information state if the protocol does not terminate. The proof will use the reverse priority order which greatly simplifies notation (in particular, the management of indices).

Lemma 6.8. *There is a unique sequence of non-redundant,* L- *and* HPW-*protecting queries on* $\mathbf{\Theta}$ *that leads to a* L- *and* HPW-*protective information state or termination.*

Proof. By Lemma 6.5, any query with a threshold lower than the highest type violates loser privacy because both a negative answer and an affirmative answer may lead to the queried agent losing. This continues to hold until $n - k - 1$ lower-priority agents have given a negative answer to a threshold query about the highest type. We will check that the $(n - k)^{\text{th}}$ query will still have the highest type as a threshold, and leads to a L- and HPW-protective information state. For the first $n - k$ queries, we only need to determine the order in which agents are queried.

Denote $h \in \{1, 2, \ldots, n\}$ the number negative answers given so far. Call agents for which a type lower than the running price is not ruled out "remaining". We show by induction over h the following statement:

$$\text{The only non-redundant query that does not violate loser or high-priority winner privacy is to the } (n-k-h)^{\text{th}} \text{ remaining agent in reverse priority order.} \tag{6.7}$$

For the rest of this proof, we will index agents in reverse priority order, meaning that agent 1 is the lowest-priority agents, and agent n is the highest-priority agent.

Note that all queries to agents who have been identified to be winners are redundant and it is without loss to abstract from them by Lemma 6.4. Also note that all queries to any remaining agent $i \leq n - k - h - 1$ make it possible that the queried agent is a loser both in the case of an affirmative and a negative answer, hence violating L privacy by

Lemma 6.5. All queries to remaining agents $i \geq n-k-h+1$ make it possible that the queried agent is a high-priority winner in the case of an affirmative or a negative answer, hence violating HPW privacy. Asking the $(n-k-h)^{\text{th}}$ remaining agent does not lead to a W or LPL contextual privacy violation.[18]

This determines the order of the first $n-k-1$ queries. Observe that agents $2, 3, \ldots, m$ for some $m \leq n$ have been asked by the property (6.7). If we show that agent 1 is the only agent that can be queried, and must be queried with a threshold that is the highest type, then the resulting information of the designer is a L- and HPW-protective information state (unless the protocol can terminate before this is the case).

We distinguish three cases. First, asking agents $1, \ldots, m$ for a lower threshold might lead to them being a loser for both an affirmative and a negative answer rendering these queries forbidden by Lemma 6.5. Second, asking agents $2, \ldots, m$ for any threshold might lead them to being a high-priority winner for both an affirmative or a negative answer, rendering these forbidden by Lemma 6.5. Third, asking agent 1 for any threshold lower than the highest type might lead to them being a loser for both an affirmative and a negative answer, making this query forbidden, also by Lemma 6.5. Asking agent 1 a threshold query with the highest type does not lead to L and HPW violations. Note that asking agent 1 complies with the agent order in (6.7). Hence, (6.7) can be seen as the definition of the overdescending-join protocol. The overdescending-join protocol for the example in the main text (with type profile $\boldsymbol{\theta} = (4, 3, 8, 2)$ and for $k = 1$ *i.e.* a second-price auction rule ϕ^{SPA}) is shown in Figure 6.12.

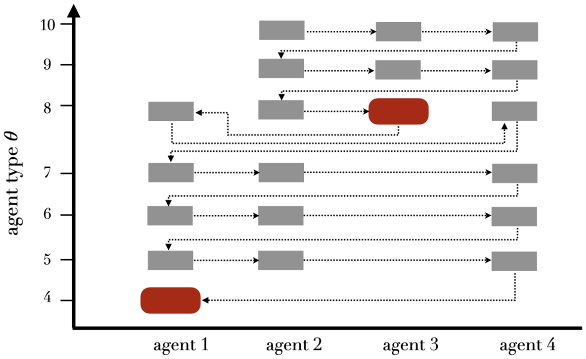


Figure 6.12: Example—Overdescending-join protocol for $\boldsymbol{\theta} = (4, 3, 8, 2)$ and ϕ^{SPA}. Queries to an agent are represented as rectangles, positioned at the corresponding threshold $\tilde{\theta}$ (gray solid rectangles represent negative responses, red rounded rectangles represent affirmative responses). The path of the dotted line represents the order in which queries are asked. Theorem 6.3 shows that the overdescending-join protocol is $\mathfrak{S}_{\text{IE}}$-maximally contextually private

[18] It may help to translate temporarily to the non-reversed priority order and consider the concrete example of the first query. We may not ask agent $n, n-1, \ldots, k+2$ as they might lose both with an affirmative and a negative answer. We may also not ask agents $k, k-1, \ldots, 1$, as they might be high-priority winner in case of a positive and a negative answer. Hence, agent $k+1$ must be queried. What's important is that an affirmative answer from the $(k+1)^{\text{st}}$ unambiguously means that they are not a loser. A negative answer from this agent means unambiguously that they are not a high-priority winner.

There are two possible cases. The first case is one in which the designer reaches information state $\mathbf{\Theta}_v$ in which $n - k$ agents have given a negative answer. In this case, the designer's information state $\mathbf{\Theta}_v$ is L- and HPW-protective. The second case is one in which at least $k + 1$ agents have given affirmative answers. In this case, the protocol may terminate (allocating to the k highest priority such agents at the $(k + 1)^{\text{st}}$ price).

□

Lemma 6.9. *On a* L- *and* HPW-*protective set* $\mathbf{\Theta}_v$, *there is a unique sequence of non-redundant,* L- *and* HPW-*protecting queries, that leads to a* L- *and* HPW-*protective set or termination.*

Proof. We reduce this to Lemma 6.8. Observe that any query to a winning agent is redundant. We may therefore drop them from consideration. Also observe that any query with a threshold higher than the running price is either trivial (if it is to a active agent), redundant (if it is to a winner) or could lead to the queried agent being a high-priority winner for both an affirmative and a negative answer (if it is to an idle agent). Hence, we may reduce to a situation in which the designer faces only active agents and needs to compute the choice rule $\phi^{(k-b)\text{-PA}}$ where b is the number of newly identified winners in the last round, on a restricted type space $\Theta' := \{\theta' \mid \theta' \leq \tilde{\theta}\}$, where $\tilde{\theta}$ is the current running price. We apply Lemma 6.8, given this choice rule and restricted type space. □

This concludes the proof. □

6.A.10 Proof of Proposition 6.8.

Proposition 6.8. *The ascending-join protocol and the overdescending-join protocol are incomparable in the contextual privacy order.*

Proof. Consider a type profile $\boldsymbol{\theta}$ where $\theta_i = \theta$ for some type $\theta \in \Theta$. In this case, by the properties of the protocols identified in (6.6) and (6.7),

$$(1, \boldsymbol{\theta}) \in \Gamma(P^{\text{AJ}}, \phi^{\text{SPA}}) \setminus \Gamma(P^{\text{OJ}}, \phi^{\text{SPA}})$$

and

$$(n, \boldsymbol{\theta}) \in \Gamma(P^{\text{OJ}}, \phi^{\text{SPA}}) \setminus \Gamma(P^{\text{AJ}}, \phi^{\text{SPA}}),$$

where P^{AJ} and P^{OJ} are the ascending and the overdescending-join protocols. □

6.A.11 Proof of Proposition 6.9.

Proposition 6.9. *The ascending-join protocol implements the second-price auction choice rule* ϕ^{SPA} *in obviously dominant strategies.*

Proof. The protocol presented in Proposition 6.9 is a *personal-clock auction* in the sense of Li (2017). The "personal clocks" in the ascending-join protocol raise running prices for agents that are active. Given this, Li (2017, Theorem 3), implies that there is an equilibrium in obviously dominant strategies. □

6.A.12 Proof of Proposition 6.10.

Proposition 6.10. *The overdescending-join protocol implements the choice rule ϕ^{SPA} in dominant strategies.*

Proposition 6.12. *The overdescending-join protocol implements the second price auction choice rule ϕ^{SPA} in dominant strategies.*

Proof. There are two types of nodes. Those in which the winner has not yet been determined, and those in which she has been determined. In the latter, all agents are indifferent between all messages, implying the incentive guarantee. In the former, before the winner has been determined, the agent does not affect the price they will be paying for any other fixed set of messages the other agents will send, leading to no profitable deviation from dropping out at an agent's type. □

6.B Two Contextually Private Protocols

We present two examples of individual elicitation protocols for common social choice functions that do not produce any contextual privacy violations. These protocols use message space $M = \Theta$, hence queries can be identified with partitions of an agent's type space.

6.B.1 Serial Dictatorships.

Recall the standard object assignment setting, in which agents may receive at most one object $c \in C$. Agents have ordinal preferences over objects, which are private information. So agents' types $\theta \in \Theta$ are preference orders of C where $\succeq_i$ reference to agent i's preference ordering. Outcomes are assignments of objects to agents $\mu \in X = 2^{N \times C}$.

Let $\mu \subseteq N \times C$ be an outcome. A *partial assignment* $N(\mu)$ is the set of agents who have an assigned object in μ, *i.e.* $N(\mu) = \{i \in N \colon \exists c \in C : (i, c) \in \mu\} \subseteq N$. If $N(\mu) = N$, we call A *complete*. For a partial assignment μ, denote $\mu(i)$ the (at most one) object assigned to agent i.

The *remaining objects* $R(\mu)$ are the objects that do not have an assigned agent in μ, *i.e.* $R(\mu) := \{c \in C \colon \nexists i \in N : (i, c) \in \mu\}$.

The designer wishes to implement the serial dictatorship assignment rule ϕ^{SD}.

To define the serial dictatorship protocol P^{SD} we characterize the nodes and edges of the rooted tree. Fix the permutation $\pi \colon N \to N$ of agents that defines the priority order of the serial dictatorship. The nodes are all partial assignments to agents in $N_i^\pi := \{\pi(i') \colon 1 \leq i' \leq i\}$ for any $i \in N$. Edges are between partial assignments μ, μ' such that exactly agents N_i^π resp. N_{i+1}^π are assigned an object, $N(\mu) = N_i^\pi$ and $N(\mu') = N_{i+1}^\pi$, and $\pi(1), \pi(2) \ldots, \pi(i)$ are assigned the same objects. We define sets of type profiles associated to each node recursively. For an edge (μ, μ'),

$$\boldsymbol{\Theta}_{\mu'} = \boldsymbol{\Theta}_\mu \cap \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \max_{\theta_{\pi(i)}} R(\mu) = \mu'(\pi(i+1)) \right\}.$$

	A	**B**
A	x or x'	x
B	x'	x or x'

	A	**B**
A	x	x
B	x'	x

	A	**B**
A	x	x'
B	x	x

	A	**B**
A	x	x
B	x'	x'

	A	**B**
A	x'	x
B	x'	x

Table 6.4: Outcomes for arbitrary efficient 2-agent social choice functions (left); under an efficient choice rule which breaks ties lexicographically (ϕ^{fair}) (mid-left, middle); under a serial dictatorship ϕ^{sd} (mid-right, right)

Here, $R(\mu)$ is the set of remaining objects when it is agent $\pi(i)$'s turn in the partial order; $\max_{\theta_{\pi(i)}} R(\mu)$ is the most preferred element of $R(\mu)$ with respect to the preference order $\theta_{\pi(i)}$. If a node is reached that is a complete assignment, the protocol ends, and the complete assignment is computed.

Proposition 6.13. *The serial dictatorship protocol $P^{SD} \in \mathcal{P}_{\Im_{\text{IE}}}$ produces no contextual privacy violations for ϕ^{SD}.*[19]

Proposition 6.13. Consider $\theta_i, \theta_i' \in \Theta$ and a partial type profile for other agents $\boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}_{-i}$ such that $(\theta_i, \boldsymbol{\theta}_{-i})$ is distinguished from $(\theta_i', \boldsymbol{\theta}_{-i})$. We will show that $\phi(\theta_i, \boldsymbol{\theta}_{-i}) \neq \phi(\theta_i', \boldsymbol{\theta}_{-i})$.

Denote μ the node of distinction. By definition of individual elicitation, this must be a query to agent i. By definition of serial dictatorship, the designer's knowledge at children of μ is given by $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid \max_{\theta_{\pi(i)}} R(\mu) = c\}$ for some $c \in R(\mu)$. Hence, if $\phi(\theta_i, \boldsymbol{\theta}_{-i})$ and $\phi(\theta_i, \boldsymbol{\theta}_{-i})$ are distinguished, agent i must get a different assignment under θ_i and θ_i', hence $\phi(\theta_i, \boldsymbol{\theta}_{-i}) \neq \phi(\theta_i, \boldsymbol{\theta}_{-i})$.

□

In the case of only two agents, serial dictatorship is the unique contextually private and efficient mechanism, as the following example shows.

Example 6.2 (Contextual Privacy and Serial Dictatorships, $n = 2$). Consider an example of two agents $N = \{1, 2\}$, each of which is allocated an object A or B. The two possible outcomes are $x = \{(1, A), (2, B)\}$ and $x' = \{(1, B), (2, A)\}$.

Table 6.4 shows possible assignments under efficiency. In the upper right table cell, efficiency requires that the outcome is x. In the lower left cell, efficiency requires that the outcome is x'. In the top left and bottom right cell, where both agents have the same type, efficiency allows either x or x'.

Four different assignments remain. The first two assignments have collective pivotality but not individual pivotality and hence produce contextual privacy violations by Theorem 6.1. The other two assignments correspond to serial dictatorships with agent orderings $\pi(1) = 1, \pi(2) = 2$ resp. $\pi(1) = 2, \pi(2) = 1$.

6.B.2 First-Price Auction.

While second-price auctions lead to contextual privacy violations, the "descending" or Dutch protocol for the first-price auction rule produces no contextual privacy violations.[20]

[19] The serial dictatorship also does not produce group or individual contextual privacy violations, as discussed in Section 6.E.

[20] An analogous observation was made in the computer science literature on decentralized computation, compare Brandt and Sandholm (2005).

A descending protocol queries, for each element of the type space $\tilde{\theta}$, in decreasing order, each agent 1 to n on whether their type θ_i is above θ. (This assumes lexicographic tiebreaking. Similar protocols do not produce contextual privacy violations for other deterministic tiebreaking rules). Formalized as a protocol, this leads to a set of nodes $N \times \Theta \times \{0, 1\}$ and edges from $((i, \theta, 0)$ to $(i+1, \theta, 0))$, for $i \in N \setminus \{n\}$ and $\theta \in \Theta$. There are edges from $(n, \theta, 0)$ to $(1, \max_{\theta' < \theta} \theta', 0)$. Furthermore, there are edges $((i, \theta, 0), (i, \theta, 1))$ for all $i \in N$ and $\theta \in \Theta$ corresponding to an agent stating that they have a type at least θ, which leads to them being allocated the good. Hence, the set of terminal nodes is $\Theta \times N \times \{1\}$. The associated set of type profiles is recursively defined as

$$\Theta_{(i+1,\theta,0),i} \coloneqq \Theta_{(i,\theta,0),i} \setminus \{\theta\} \tag{6.8}$$

$$\Theta_{(i,\theta,1),i} \coloneqq \{\theta\} \tag{6.9}$$

$$\Theta_{(1,\theta,0),i} \coloneqq \Theta_{(n,\max_{\theta'<\theta} \theta',0),i} \setminus \{\theta\}, \tag{6.10}$$

where $\Theta_{v,i}$ denotes the projection of $\boldsymbol{\Theta}_v$ onto agent i's type. For all non-queried agents at node v, and a child w, $\Theta_{w,i} = \Theta_{v,i}$. (6.8) rules out the type θ for type i when they respond that they are not type θ. (6.9) identifies an agent's type exactly when they respond they are type θ. (6.10) rules out the type θ for agent n when they respond they are not θ and leads to the protocol considering the next-lowest type $\max\{\theta' \mid \theta' < \theta\}$. This defines a protocol $P^{\text{desc}} \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$.

Proposition 6.14. *The descending protocol $P^{desc} \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$ for the first-price rule ϕ^{FP} does not produce any contextual privacy violations.*

Proof of Proposition 6.14. First note that by construction of the protocol, the first query leading to a singleton possible type space must be a type θ_i attaining $\max_{i \in N} \theta_i$. This implies that the descending protocol is a protocol for the first-price choice rule (with tie-breaking according to the order $1, 2, \dots, n$).

To show that P^{desc} does not lead to contextual privacy violations, let v be a node (querying agent i) that distinguishes $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$. Note that terminal nodes cannot distinguish type profiles. Hence, $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$ are distinguished at a node of the form $(i, \tilde{\theta}, 0)$. By definition of the descending protocol, the children of the node $(i, \tilde{\theta}, 0)$ are associated to sets

$$\{\tilde{\theta}\} \text{ and } \{\tilde{\theta}' \in \Theta \colon \tilde{\theta}' < \tilde{\theta}\}.$$

Let, without loss, $\theta_i = \tilde{\theta}$ and $\theta_i' < \tilde{\theta}$. In the former case, the outcome is that agent i gets the good at price $\tilde{\theta}$. By definition of the descending protocol, in the latter case, it is that either agent i does not get the good, or they get it at a price $\tilde{\theta}' < \tilde{\theta}$. In particular, $\phi(\theta_i, \boldsymbol{\theta}_{-i}) \neq \phi(\theta_i', \boldsymbol{\theta}_{-i})$.

□

6.C Other Maximally Contextually Private Protocols

In Section 6.4, we derived two maximally contextual private choice rules for the second-price auction rule. In particular, Theorem 6.3 showed that the ascending-join and overdescending-join protocols are maximally contextually private for the second-price auction rule. We also argued that these protocols give general principles for understanding maximally contextually private protocols: they choose a specific set of agents to protect, and then delay asking questions to those agents.

In general, there will be other ways to get maximally contextually private protocols, through "guessing." That is, the designer may eliminate *all* contextual privacy violations at some type profile by simply verifying a correct "guess" of the relevant statistics of the type profile. For example, the "guessing" protocol for the first-price auction rule is essentially a posted price, and, when correct, it avoids contextual privacy violations altogether.

We illustrate this principle again through the second-price auction rule. Here, we note that the ascending-join and overdescending-join protocols are not the unique maximally contextually private protocols for the second-price auction rule. To show the existence of other maximally contextually private protocols, we need only note that there are protocols that are incomparable in the contextual privacy order to these two. The following example gives on such protocol, the *guessing protocol*.

The *guessing protocol* for the second-price choice rule "guesses" a price $\tilde{\theta} \in \Theta$ and an agent who has exactly this value, and then aims to verify the guess. The protocol $P^{\tilde{\theta}\text{-guess}}$ starts with a threshold query for type $\tilde{\theta}$ for all agents. If all but one agent answer negatively, the protocol asks the agent who the designer guesses has a value of exactly $\tilde{\theta}$ whether this is indeed the case.[21]

Consider again the type profile from the example in Section 6.4.2. There are 4 agents, and the type profile is $\boldsymbol{\theta}^* = (4, 3, 8, 2)$. Let's consider first the contextual privacy violations of a guessing protocol with a guess of $\tilde{\theta} = 4$, and $\{j \mid \theta_j = \boldsymbol{\theta}^*_{[2]}\} = \{1\}$. That is, the designer guesses that the second-highest bid is agent 1's, and that it is 4. The first threshold query in the protocol $P^{4\text{-guess}}$ is for the threshold 5. This protocol begins by asking every agent if their type is strictly greater than 5, beginning with agent 1. All agents except agent 3 answer negatively. Then, the protocol asks only agent 1 if her type is strictly above 4. Agent 1 answers affirmatively. The designer gets enough information to compute the rule, and there are no contextual privacy violations for any agent. That is,

$$N \times \{\boldsymbol{\theta}^*\} \cap \Gamma(P^{\text{AJ}}, \phi^{\text{SPA}}) = \emptyset \tag{6.11}$$

Now compare (6.11) to the privacy violations of the ascending-join protocol P^{AJ} at $\boldsymbol{\theta}^*$. As we know from the main text, the ascending-join protocol leads to a violation for agent 2 at $\boldsymbol{\theta}^*$, *i.e.*

$$(2, \boldsymbol{\theta}^*) \in \Gamma(P^{\text{AJ}}, \phi^{\text{SPA}}). \tag{6.12}$$

Similarly, the overdescending-join protocol P^{OJ} leads to a violation for the winner, agent 3. That is,

$$(3, \boldsymbol{\theta}^*) \in \Gamma(P^{\text{OJ}}, \phi^{\text{SPA}}). \tag{6.13}$$

So, we see that *at this particular type profile* $\boldsymbol{\theta}^*$, the violations from the guessing protocol form a subset of the violations of both the ascending-join and overdescending-join protocol.

[21] Otherwise, if the guess is not correct, the protocol proceeds in any manner compatible with bimonotonicity—or the point we make here, it will not matter exactly how the protocol proceeds when the guess is wrong.

As P^{AJ} and P^{OJ} are maximally contextually private, it must be that $P^{4-\text{guess}}$ has contextual privacy violations that neither of these protocols have. In particular, there is a maximally contextually private protocol with weakly fewer violations than $P^{4-\text{guess}}$ whose contextual privacy violations are incomparable to those of the ascending-join and overdescending-join protocols.[22]

The guessing protocol suggests an interesting direction for future work: If the designer cares about the *expected* contextual privacy violations, then the guessing protocol may be preferable to the ascending-join or overdescending-join protocols. In other words, if the designer uses their prior $f(\theta)$ to choose a "good guess", and if the designer cares about privacy violations *in expectation*, then the guessing protocol might be preferable, from a privacy standpoint, to the ascending or overdescending protocols, because the designer can leverage their prior to minimize contextual privacy violations in expectation. But, if the designer can't use their prior to choose a "good guess", because *e.g.*, the prior is diffuse, then the guessing protocol may produce more contextual privacy violations in expectation than the ascending or overdescending protocols.

6.D The Count Elicitation Technology

The framework laid out in Section 6.2 captures a wide range of possible environments via the elicitation technology. Most of the paper focuses on a particular minimal trust elicitation technology, the individual elicitation technology. Here we define another elicitation technology that corresponds to a similarly simple assumption—it corresponds to an ability for the designer to *anonymize* messages, or otherwise commit to forget the labels of the messages received.

The *count elicitation technology* $\mathfrak{S}_{\text{Count}}$ is able to provide counts of all messages sent, and consists of a single partition

$$\mathfrak{S}_{\text{Count}} = \{\{\boldsymbol{b} \in \mathbf{M} \colon |\{m_i = m\}| = k\} : m \in M, k \in N\}$$

One example for a binary message space $M = \{0, 1\}$ is a ballot box: Agents that are sending message 1 are consider to vote *for* an issue, those sending message 0 *against*.

[22] To see this in the example considered in the main text, consider another particular type profile $\boldsymbol{\theta}' = (6, 3, 8, 2)$. At this type profile, the violations for the ascending-join and over-descending join protocol remain the same as in (6.12) and (6.13), respectively (replacing $\boldsymbol{\theta}^*$ with $\boldsymbol{\theta}'$). However the 4-guessing protocol $P^{\text{4-guess}}$ must produce at least a violation for agent 1, *i.e.*

$$(1, \boldsymbol{\theta}') \in \Gamma(P^{\text{4-guess}}, \phi^{\text{SPA}}). \tag{6.14}$$

Given this, we know that it is the case that

$$\Gamma(P^{\text{AJ}}, \phi^{\text{SPA}}) \not\subset \Gamma(P^{\text{4-guess}}, \phi^{\text{SPA}}) \text{ and } \Gamma(P^{\text{4-guess}}, \phi^{\text{SPA}}) \not\subset \Gamma(P^{\text{AJ}}, \phi^{\text{SPA}}),$$

and that

$$\Gamma(P^{\text{OJ}}, \phi^{\text{SPA}}) \not\subset \Gamma(P^{\text{4-guess}}, \phi^{\text{SPA}}) \text{ and } \Gamma(P^{\text{4-guess}}, \phi^{\text{SPA}}) \not\subset \Gamma(P^{\text{OJ}}, \phi^{\text{SPA}}).$$

So, we have shown that there is a protocol that is incomparable to both the ascending-join and the overdescending-join protocols. This implies that there must be some maximally contextually private protocol that is neither the ascending-join nor the overdescending-join protocol, as there must be a protocol that includes the violation $(1, \boldsymbol{\theta}')$ that is also maximal.

The count elicitation technology requires more trust than the individual elicitation technology in the sense that it requires "mediation". To see this, suppose the designer uses the count elicitation technology to ask "How many agents have a message above x?" Then, in order for it to be the case that the designer *only* learns the number, and not the identity of the agents, there must be a technology that anonymizes agents messages when they are sent. In practice, count elicitation technologies could be a ballot box, a third-party mediator, or another trusted anonymization technique.

Our techniques help us to understand that the count elicitation technology has similar properties as the individual elicitation technology. A first statement shows that for any $\mathfrak{S}_{\text{IE}}$-protocol P there is a $\mathfrak{S}_{\text{Count}}$-protocol P' with the same contextual privacy violations.

Proposition 6.15. *Let $(P, \boldsymbol{\sigma}) \in \mathcal{P}_{\mathfrak{S}_{IE}}$ be a protocol for ϕ. Then, there is a protocol $(P', \boldsymbol{\sigma}') \in \mathcal{P}_{\mathfrak{S}_{\text{Count}}}$ for ϕ such that $\Gamma(P', \boldsymbol{\sigma}', \phi) \subseteq \Gamma(P, \boldsymbol{\sigma}, \phi)$.*

Proof. If the message space contains only one element, then only constant social choice functions can be computed, in which case, the result is true.

Otherwise, consider any query in a protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{IE}}}$. We construct P' by induction on the tree P. By our induction hypothesis, we may assume that the designer's knowledge at node v in protocol P' is the same as it is at node v in protocol P. For the induction step, we choose a $\mathfrak{S}_{\text{Count}}$-query s'_v and a strategy profile $\boldsymbol{\sigma}$ such that the designer also has the same knowledge at every child of v. By definition of $\mathfrak{S}_{\text{IE}}$, there must be an agent i such that the child reached only depends on the message sent by agent i. Choose any $m \in M$. i keeps their messaging strategy, all other agents send message m. The protocol determines the outcome based on the observed counts: $n - 1$ agents are sending message m in any outcome, so the message of agent i can be determined, and the child can be computed.

The so constructed protocol has the same set of distinguished types. Hence, it is still a protocol for ϕ and has the same contextual privacy violations. □

This result may appear somewhat unnatural. The construction of the protocol P' strongly relies on the designer's ability to choose agents' strategies $\boldsymbol{\sigma}'$. Any deviation from these strategies $\boldsymbol{\sigma}'$ makes the "decoding" of a single agent's action impossible. Hence, in the presence of strategic incentives (where $\boldsymbol{\sigma}$ is not determined by the designer), this reduction will be of little use.

One might wonder whether count is then strictly more powerful for the purposes of contextual privacy. Interestingly, the count elicitation technology also fails in cases of collective and individual pivotality.

Proposition 6.16. *Let ϕ be a social choice function, and consider a particular type profile $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. If for any subset $A \subseteq N$ of agents, and types $\theta'_i \in \Theta$ for agents $i \in A$, if*

$$\phi(\boldsymbol{\theta}) \neq \phi(\boldsymbol{\theta}'_A, \boldsymbol{\theta}_{-A}) \qquad \text{(collective pivotality)}$$

and for all $i \in A$

$$\phi(\boldsymbol{\theta}) = \phi(\theta'_i, \boldsymbol{\theta}_{-i}) \qquad \text{(no individual pivotality)}$$

then for any count elicitation protocol $P \in \mathcal{P}_{\mathfrak{S}_{\text{Count}}}$, there exists an agent $i \in A$ whose contextual privacy is violated at $\boldsymbol{\theta}$, i.e. *$(i, \boldsymbol{\theta}) \in \Gamma(P, \phi)$.*

Proof. Consider the first query v that distinguishes types in the set

$$\tilde{\mathbf{\Theta}} = \{\boldsymbol{\theta}_{-A}\} \times \bigtimes_{i \in A} \{\theta_i, \theta_i'\}.$$

This query must exist because of collective pivotality. By the definition of individual elicitation, there must be an agent i that sends different messages for θ_i and θ_i', $\sigma_i(\theta, v) \neq \sigma_i(\theta', v)$. By the lack of individual pivotality, observe that $\phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i})$. By the definition of the count elicitation technology, $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$ are distinguished at v. □

As a consequence of Proposition 6.16, all of the results on collective and individual pivotality obtained in Section 6.3.2 also hold for the count technology, which can guide intuitions even for this richer elicitation technology.

6.E Variations of Contextual Privacy

In this section, we consider two ways to strengthen our definition of contextual privacy violations. The two key pieces of the definition of contextual privacy violations are whether two types for an agent are distinguished and whether there is some difference to the outcome that would justify the distinction. The following two variations of contextual privacy violations consider stronger notions of distinction and justification, respectively.

We explore these stronger concepts for both theoretical and practical reasons. On the practical side, these extensions may map onto design goals in some settings. On the theoretical side, these criteria help to illuminate connections to other desiderata in mechanism design, and illustrate which of our results are robust to alternative formulations of contextual privacy.

6.E.1 Individual Contextual Privacy

The first extension, the *individual contextual privacy violation*, requires that if two types are distinguishable for agent i, these types must lead to different outcomes under ϕ *for agent* i. This definition thus only applies in domains where the outcome space X, specifies an allocation for each agent $i \in N$. Let $\phi_i(\boldsymbol{\theta})$ denote the outcome under ϕ for agent i.

Definition 6.17 (Individual Contextual Privacy Violations)**.** Let $(P, \boldsymbol{\sigma}) \in \mathcal{P}_{\mathfrak{S}}$ be a protocol for ϕ. Then the set of *individual contextual privacy violations* $\Gamma_{\text{ind.}}(P, \boldsymbol{\sigma}, \phi) \subseteq N \times \mathbf{\Theta}$ contains tuples $(i, \boldsymbol{\theta})$ for which there exists $\theta_i' \in \Theta$ such that

$$P \text{ distinguishes } (\theta_i, \boldsymbol{\theta}_{-i}) \text{ and } (\theta_i', \boldsymbol{\theta}_{-i}) \text{ yet } \phi_i(\theta_i, \boldsymbol{\theta}_{-i}) \neq \phi_i(\theta_i', \boldsymbol{\theta}_{-i}).$$

We call a choice rule *individually contextually private* for ϕ if there is a protocol $(P, \boldsymbol{\sigma})$ such that $\Gamma_{\text{ind.}}(P, \boldsymbol{\sigma}, \phi) = \emptyset$.

As for $\Gamma(P, \boldsymbol{\sigma}, \phi)$, we may also omit $\boldsymbol{\sigma}$ from our notation if the strategies are clear from context. Notice that individual contextual privacy is stronger than contextual privacy—any choice rule that is individually contextually private is also contextually private. If there were an agent i for whom contextual privacy were violated, then individually contextual privacy would automatically be violated.

As a normative criterion, individual contextual privacy requires that if the designer can distinguish between two types for agent i, then it *should* be the case that agent i's outcome is changed. This criterion captures a notion of legitimacy—agent i may view participation in the mechanism as involving an inherent tradeoff between information revelation and allocation. We can imagine a speech from agent i along the following lines: "The designer can learn that I have type θ_i and not θ_i' as long as the designer's knowledge of this makes a difference to my allocation."

Individual contextual privacy is closely related to *non-bossiness*, introduced by Satterthwaite and Sonnenschein (1981). A choice rule ϕ is *non-bossy* if for all $\theta_i, \theta_i' \in \Theta, \boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}_{-i}$,

$$\phi_i(\theta_i, \boldsymbol{\theta}_{-i}) = \phi_i(\theta_i', \boldsymbol{\theta}_{-i}) \implies \phi(\theta_i, \boldsymbol{\theta}_{-i}) = \phi(\theta_i', \boldsymbol{\theta}_{-i}). \tag{6.15}$$

We define the set of *non-bossiness violations* $B(P, \phi) \subseteq N \times \boldsymbol{\Theta}$ consisting of $(i, \boldsymbol{\theta})$ where (6.15) fails to hold.

Non-bossiness says that if agent i changes her report from θ_i to θ_i' and her allocation is unchanged, then no other agent j's allocation changes either. The idea is that if agent i could unilaterally change her report and affect a change in some agent j's allocation without changing her own allocation, agent i would be "bossy."

We can characterize the set of individual contextual privacy violations $\Gamma_{\text{ind.}}$ in terms of contextual privacy violations and bossiness violations.

Proposition 6.17. *The set of individual contextual privacy violations is the intersection of contextual privacy violations and non-bossiness violations, $\Gamma_{ind.}(P, \phi) = \Gamma(P, \phi) \cup B(P, \phi)$.*

That is, the set of individual contextual privacy violations is the union of contextual privacy and non-bossiness violations.

Proof. We first show that $B(P, \phi) \subseteq \Gamma_{\text{ind.}}(P, \phi)$. Let $(i, \boldsymbol{\theta})$ be a non-bossiness violation. Hence, there exists a $j \in N \setminus \{i\}$ and type profiles $(\theta_i, \boldsymbol{\theta}_{-i})$, $(\theta_i', \boldsymbol{\theta}_{-i})$ such that $\phi_i(\theta_i, \boldsymbol{\theta}_{-i}) = \phi_i(\theta_i', \boldsymbol{\theta}_{-i})$ but $\phi_j(\theta_i, \boldsymbol{\theta}_{-i}) \neq \phi_j(\theta_i', \boldsymbol{\theta}_{-i})$. Because of the latter, the protocol P must distinguish $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$. By the former property, this leads to an individual contextual privacy violation for agent i. By definition, $\Gamma(P, \phi) \subseteq \Gamma_{\text{ind.}}(P, \phi)$. Hence, $B(P, \phi) \cup \Gamma(P, \phi) \subseteq \Gamma_{\text{ind.}}(P, \phi)$.

To show the converse, we show that $(N \times \boldsymbol{\Theta}) \setminus (\Gamma(P, \phi) \cup B(P, \phi)) = ((N \times \boldsymbol{\Theta}) \setminus \Gamma(P, \phi)) \cap ((N \times \boldsymbol{\Theta}) \setminus B(P, \phi)) \subseteq (N \times \boldsymbol{\Theta}) \setminus \Gamma_{\text{ind.}}(P, \phi)$, which finishes the proof. In other words, we show that if at $(i, \boldsymbol{\theta})$ there is neither a contextual privacy nor a non-bossiness violation, there can't be an individual contextual privacy violation. Consider any $(\theta_i, \boldsymbol{\theta}_{-i})$ and $(\theta_i', \boldsymbol{\theta}_{-i})$ that P distinguishes. By contextual privacy, $\phi(\theta_i, \boldsymbol{\theta}_{-i}) \neq \phi(\theta_i', \boldsymbol{\theta}_{-i})$. By non-bossiness, $\phi_i(\theta_i, \boldsymbol{\theta}_{-i}) \neq \phi_i(\theta_i', \boldsymbol{\theta}_{-i})$ follows. Thus, P does not produce an individual contextual privacy violation at $(i, \boldsymbol{\theta})$. □

Given this relationship of individual contextual privacy and non-bossiness, we are able to leverage existing results to yield unique characterizations for the first-price auction and the serial dictatorship.

We say a mechanism is *neutral* if for all type profiles $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and permutations π, $\phi(\pi\boldsymbol{\theta}) = \pi\phi(\boldsymbol{\theta})$.

Proposition 6.18. *An object assignment choice rule ϕ is individually contextually private, neutral and strategyproof if and only if it is a serial dictatorship.*

Proof. The serial dictatorship is contextually private and non-bossy, hence individually contextually private by Proposition 6.17. It also is strategyproof.

Conversely, if ϕ is individually contextually private, then it is also non-bossy by Proposition 6.17. By Svensson (1999), a mechanism is neutral, strategyproof and non-bossy only if it is a serial dictatorship. So, if ϕ is individually contextually private, neutral and strategyproof, then it is a serial dictatorship. □

Note here that efficiency is implied by neutrality. A characterization of efficient, strategyproof and non-bossy mechanisms is supplied by Pycia and Ünver (2017) and Bade (2020)—note that the class characterized in those works is larger than just serial dictatorships.

Similarly, we can draw on existing literature on non-bossiness to uniquely characterize the first-price auction.

Proposition 6.19. *The first-price auction is the unique individually contextually private, efficient, and individually rational auction rule.*

Proof. The protocol for the first-price auction rule presented in Section 6.B.2 is individually contextually private. It is well known that the first-price auction is efficient and individually rational.

Let ϕ be any individually contextually private, efficient and individually rational auction rule. By the characterization of efficient, non-bossy and individually rational auctions from Pycia and Raghavan (2021, Theorem 1), this means that it is a protocol for the first-price auction. □

6.E.2 Group Contextual Privacy

Another extension is *group* contextual privacy violations, which strengthens the notion of distinction. Group contextual privacy requires that if a protocol distinguishes two type *profiles*, they must lead to different outcomes. This notion strengthens contextual privacy, which requires only that if a single agent's types are distinguishable, then they lead to different outcomes.

Definition 6.18 (Group Contextual Privacy Violations)**.** A protocol $P = (V, E)$ for a social choice function ϕ with strategies $\boldsymbol{\sigma}$ produces a *group contextual privacy violation* at $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ if there is a type $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ such that

$$P \text{ distinguishes } \boldsymbol{\theta} \text{ and } \boldsymbol{\theta}' \text{ yet } \phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}').$$

Notice again that this definition strengthens contextual privacy. Here, it is because it strengthens the underlying notion of distinguishability—two type profiles $\boldsymbol{\theta}, \boldsymbol{\theta}'$ are distinguishable if they belong to different terminal nodes. Regular contextual privacy's notion of distinguishability is on the agent-level—two types θ_i, θ_i' are distinguishable if they belong to different terminal nodes, *holding all other agent's types fixed* at $\boldsymbol{\theta}_{-i}$.

We characterize the set of group contextually private protocols next. Denote the set of outcomes reachable from node v by X_v.

Theorem 6.5. *A protocol $P = (V, E)$ is $\mathfrak{S}$-group contextually private if and only if for any query, $\bigcup_{w \in \text{children}(v)} X_w = X_v$ is a disjoint union.*

Proof. First assume that P is group contextually private, and assume for contradiction that v is a query such that $(v, w), (v, w') \in E$ and $X_w \cap X_{w'} \neq \emptyset$. Hence, there are $\boldsymbol{\theta} \in \boldsymbol{\Theta}_w$ and $\boldsymbol{\theta}' \in \boldsymbol{\Theta}_{w'}$ such that $\phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}')$, which contradicts group contextual privacy.

Next assume that reachable outcomes are disjoint at each query. Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ be distinguished at v. As outcomes are disjoint, it must be that $\phi(\boldsymbol{\theta}) \neq \phi(\boldsymbol{\theta}')$. Hence the protocol is group contextually private. □

This general characterization lends more practical insight when specialized to group contextual privacy under only individual elicitation protocols.

Corollary 6.1. *A social choice function is group contextually private under individual elicitation protocols if and only if it can be represented by a protocol in which, at every node, the agent's choice rules out a subset of the outcomes.*

This characterization, in particular, implies that the serial dictatorship and the first price auction are group contextually private. In the serial dictatorship protocol, whenever an agent is called to play, they obtain their favorite object among those that remain. So, their choice rules out the outcomes in which a different agent gets their favorite object that remains. In the first price auction, the agent agent renders a particular outcome impossible, namely the outcome under which they win the good at a particular price.

Under individual elicitation protocols, group contextual privacy is thus reminiscent of other extensive-form properties related to simplicity that the serial dictatorship satisfies. In particular, the serial dictatorship is obviously strategyproof Li (2017) and strongly obviously strategyproof Pycia and Troyan (2023).

Is there are containment relationship between rules that admit group contextually private protocols and those that admit an implementation in obviously dominant strategies? It turns out that the answer is no: there are mechanisms that are group contextually private and not obviously strategyproof, and vice versa. The ascending auction is obviously strategyproof Li 2017, but not group contextually private with respect to individual elicitation (it is a protocol for the second-price auction, which is not contextually private). A class of "non-clinching rules", on the other hand, are strategyproof and group contextually private, but fail to have an obviously strategyproof implementation. We next offer an example of a non-clinching rule, and show that it is group contextually private but not obviously strategyproof.

Example 6.3 (Non-Clinching Rule). In particular, there are strategyproof choice rules that are not obviously strategyprouf but group-contextually private. As an example, consider $n = 2$, $\Theta = \{\underline{\theta}, \overline{\theta}\}$ and $X = \{x_1, x_2, x_3, x_4\}$. Assume that for agent 1,

$$x_1 \succ_{\underline{\theta}} x_3 \succ_{\underline{\theta}} x_2 \succ_{\underline{\theta}} x_4$$
$$x_1 \prec_{\underline{\theta}} x_3 \prec_{\underline{\theta}} x_2 \prec_{\underline{\theta}} x_4$$

and for agent 2

$$x_1 \succ_{\underline{\theta}} x_2 \succ_{\underline{\theta}} x_3 \succ_{\underline{\theta}} x_4$$
$$x_1 \prec_{\underline{\theta}} x_2 \prec_{\underline{\theta}} x_3 \prec_{\underline{\theta}} x_4.$$

Consider the social choice function

$$\phi(\underline{\theta}, \underline{\theta}) = x_1 \qquad \phi(\underline{\theta}, \overline{\theta}) = x_2 \qquad \phi(\overline{\theta}, \underline{\theta}) = x_3 \qquad \phi(\overline{\theta}, \overline{\theta}) = x_4.$$

As ϕ is injective, any protocol for ϕ is group contextually private. It is also tedious but straightforward to check that this rule is strategyproof. There is no obviously strategyproof implementation, however. Assume that agent 1 is asked to play first. They face a choice between outcomes $\{x_1, x_3\}$ and $\{x_2, x_4\}$, which, for both $\underline{\theta}$ and $\overline{\theta}$ types are are not ordered in the set order, and hence make no action obviously dominated. A similar observation for agent 2 shows that neither first action can be obviously dominant.

Chapter 7

Conclusion

This dissertation explored the economic engineering of personalized experiences by formalizing key algorithmic and economic concepts relating to the market impacts of personalization algorithms. We introduced the notion of algorithmic demand, characterized its dependence on preference measurement noise, search, and algorithm choice, and analyzed its implications for market concentration and consumer surplus. We formalized privacy in personalization through contextual integrity and showed how, when combined with other desiderata, privacy protection needs to be allocated like other scarce goods. Finally, we investigated alternative user-algorithm interactions, including using regret data and the design of generative models. All of these contributions contribute to scientific foundations for the economic engineering of personalized experiences.

These findings have significant implications for the design and regulation of personalization systems. For regulators, our findings in Chapters 2 to 4 identify key factors driving market concentration in algorithmic personalization: biased algorithmic estimates, measurement noise, and user search. Interventions in this area, much like investigations on algorithmic collusion (Calvano et al. 2020) hence need to pay attention to these channels, which shape algorithmic demand. As we discuss in Chapter 5, regret data could be used in future consumer protection regimes as "early-warning systems" for deceptive or manipulative designs. Finally, contextual privacy, as an alternative to differential privacy (Dwork 2006), highlights the possibility of not querying information at all to preserve privacy, hence protecting the privacy (of some), unconditionally. Designs based on contextual privacy may enable a more faithful implementation of purpose limitation in social domains.

While this work provides a theoretical foundation, several limitations suggest avenues for future research. The thesis (with the exception of a survey presented in Chapter 4) considers theoretical analysis or simulations and does not include empirical validation. Much of this relies on the challenge for researchers to get access to personalization data in the wild (compare A. Haupt, Hitzig, and Gleason (2023)). At least Chapter 5 lends itself to deployment in middleware. Other opportunities arise out of our focus on the demand side: Neither did we consider content supply, nor entry or pricing, see (Jagadeesan, N. Garg, and Steinhardt 2023). What is, for example, the result on entry of the patterns we observed in algorithmic demand? What can be said about the emergence of *virality*, so very high probability of a few personalized experiences? An additional limitation of the models presented here lies in our assumptions on preferences, which we model as latent yet fixed. This contrasts with the importance of habit formation online, and is subject of a related literature Curmei et al. (2022), Carroll et al. (2021), Allcott, Gentzkow, and Song (2022), and Becker and Murphy (1988). Taking into account habit formation, addictive qualities, and changing preferences in the study of market effects of algorithms can help. A direct question empirically is whether the selection of personalized experiences has an important role for habit formation (in comparison to design decisions beyond the selection of personalized experiences). A final limitation is that the platform we assume that the algorithm designer maximizes welfare. While many of the intuitions in the chapters presented here continue to hold under some amount of conflict of interest, the interaction of preference measurement error and concentration in Chapter 4 could get richer when a conflict of interest is considered, and allows to connect to studies of biased intermediation, *e.g.*, on the Amazon Marketplace Cornière and G. Taylor (2019), K. H. Lee and Musolff (2021), Gutierrez Gallardo (2021), and Hartzell and A. Haupt (2025).

There are opportunities for future research informed by our findings and methodological advances. Chapter 2's intuition for why algorithms may not favor outcomes with noisily observed feedback may generalize to gradient-based algorithms. Higher amounts of noise lead to "diffusion" from areas with noisy (gradient) feedback, allowing to understand inferential bias in more general learning, compare Damian, Ma, and J. D. Lee (2021). Chapter 4 suggests the study of *symmetric persuasion* in more general settings, as a novel constraint enforcing symmetric informativeness. Chapter 3 promises a much richer theory of mechanisms when costs are allowed to be random and correlated with types, and under a conflict of interest between the designer and the agents, such as under revenue maximization.

As machine learning permeates more and more areas of human existence, the effects of algorithm design will only increase. As personalization is crucial for *useful* machine learning systems, we expect the market effects of personalization algorithms to only increase, promising an increasing role for the economic engineering of personalized experiences.

References

Abdollahpouri, Himan (2019). “Popularity Bias in Ranking and Recommendation”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’19. Honolulu, HI, USA: Association for Computing Machinery, pp. 529–530. ISBN: 9781450363242. DOI: 10.1145/3306618.3314309. URL: https://doi.org/10.1145/3306618.3314309.

Abdulkadiroğlu, Atila, Parag A Pathak, and Alvin E Roth (Apr. 2005). “The New York City High School Match”. In: *American Economic Review* 95.2, pp. 364–367. ISSN: 0002-8282. DOI: 10.1257/000282805774670167. URL: http://dx.doi.org/10.1257/000282805774670167.

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar (Nov. 2022). “Too Much Data: Prices and Inefficiencies in Data Markets”. In: *American Economic Journal: Microeconomics* 14.4, pp. 218–256. ISSN: 1945-7685. DOI: 10.1257/mic.20200200. URL: http://dx.doi.org/10.1257/mic.20200200.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman (June 2016). “The Economics of Privacy”. In: *Journal of Economic Literature* 54.2, pp. 442–492. ISSN: 0022-0515. DOI: 10.1257/jel.54.2.442. URL: http://dx.doi.org/10.1257/jel.54.2.442.

Agan, Amanda, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan (Feb. 2023). *Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias*. Tech. rep. National Bureau of Economic Research. DOI: 10.3386/w30981. URL: http://dx.doi.org/10.3386/w30981.

Ahmadi, Saba, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita (2021). “The Strategic Perceptron”. In: *Proceedings of the 22nd ACM Conference on Economics and Computation*. EC ’21. Budapest, Hungary: Association for Computing Machinery, pp. 6–25. ISBN: 9781450385541. DOI: 10.1145/3465456.3467629. URL: https://doi.org/10.1145/3465456.3467629.

Akbarpour, Mohammad and Shengwu Li (2020). “Credible Auctions: A Trilemma”. In: *Econometrica* 88.2, pp. 425–467. DOI: https://doi.org/10.3982/ECTA15925. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15925. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15925.

Alabdulrahman, Rabaa and Herna Viktor (2021). “Catering for unique tastes: Targeting grey-sheep users recommender systems through one-class machine learning”. In: *Expert Systems with Applications* 166, p. 114061. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.114061. URL: https://www.sciencedirect.com/science/article/pii/S0957417420308241.

Alawadhi, Neha (2021). “Only 1 in 250 understand role of encryption in securing messaging: Study”. In: *Business Standard* March 21.

Ali, S Nageeb, Greg Lewis, and Shoshana Vasserman (July 2022). “Voluntary Disclosure and Personalized Pricing”. In: *The Review of Economic Studies* 90.2, pp. 538–571. ISSN: 0034-6527. DOI: 10.1093/restud/rdac033. eprint: https://academic.oup.com/restud/article-pdf/90/2/538/49551460/rdac033.pdf. URL: https://doi.org/10.1093/restud/rdac033.

Allcott, Hunt, Matthew Gentzkow, and Lena Song (July 2022). “Digital Addiction”. In: *American Economic Review* 112.7, pp. 2424–63. DOI: 10.1257/aer.20210867. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20210867.

Altonji, Joseph G. and Charles R. Pierret (2001). “Employer Learning and Statistical Discrimination”. In: *The Quarterly Journal of Economics* 116.1, pp. 313–350. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/2696451 (visited on 01/31/2025).

Alvarez, Ramiro and Mehrdad Nojoumian (2020). “Comprehensive survey on privacy-preserving protocols for sealed-bid auctions”. In: *Computers & Security* 88, p. 101502. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2019.03.023. URL: https://www.sciencedirect.com/science/article/pii/S0167404818306631.

Anderson, Chris (July 2006). *The long tail*. Hachette Books.

Angrist, J D and Jorn-Steffen Pischke (Dec. 2008). *Mostly harmless econometrics*. en. Princeton, NJ: Princeton University Press.

Ashlagi, Itai and Yannai A. Gonczarowski (2018). “Stable matching mechanisms are not obviously strategy-proof”. In: *Journal of Economic Theory* 177, pp. 405–425. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2018.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0022053118303454.

Auer, Peter (Mar. 2003). “Using confidence bounds for exploitation-exploration trade-offs”. In: *J. Mach. Learn. Res.* 3.null, pp. 397–422. ISSN: 1532-4435.

Auer, Peter, Nicolò Cesa-Bianchi, and Paul Fischer (2002). “Finite-time Analysis of the Multiarmed Bandit Problem”. In: *Machine Learning* 47.2/3, pp. 235–256. ISSN: 0885-6125. DOI: 10.1023/a:1013689704352. URL: http://dx.doi.org/10.1023/A:1013689704352.

Ausubel, Lawrence M. (Dec. 2004). “An Efficient Ascending-Bid Auction for Multiple Objects”. In: *American Economic Review* 94.5, pp. 1452–1475. DOI: 10.1257/0002828043052330. URL: https://www.aeaweb.org/articles?id=10.1257/0002828043052330.

Bade, Sophie (2020). “Random Serial Dictatorship: The One and Only”. In: *Mathematics of Operations Research* 45.1, pp. 353–368. DOI: 10.1287/moor.2019.0987. eprint: https://doi.org/10.1287/moor.2019.0987. URL: https://doi.org/10.1287/moor.2019.0987.

Bade, Sophie and Yannai A. Gonczarowski (2017). “Gibbard-Satterthwaite Success Stories and Obvious Strategyproofness”. In: EC ’17, p. 565. DOI: 10.1145/3033274.3085104. URL: https://doi.org/10.1145/3033274.3085104.

Baek, Jackie and Vivek Farias (2021). “Fair Exploration via Axiomatic Bargaining”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 22034–22045. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/b90c46963248e6d7aab1e0f429743ca0-Paper.pdf.

Baek, Jackie and Ali Makhdoumi (2023). *The Feedback Loop of Statistical Discrimination*. Tech. rep. DOI: 10.2139/ssrn.4658797. URL: http://dx.doi.org/10.2139/ssrn.4658797.

Balinski, Michel and Tayfun Sönmez (1999). “A Tale of Two Mechanisms: Student Placement”. In: *Journal of Economic Theory* 84.1, pp. 73–94. ISSN: 0022-0531. DOI: https://doi.

org/10.1006/jeth.1998.2469. URL: https://www.sciencedirect.com/science/article/pii/S0022053198924693.

Ball, Ian (Feb. 2025). “Scoring Strategic Agents”. In: *American Economic Journal: Microeconomics* 17.1, pp. 97–129. DOI: 10.1257/mic.20230275. URL: https://www.aeaweb.org/articles?id=10.1257/mic.20230275.

Banchio, Martino and Giacomo Mantegazza (2023). “Adaptive Algorithms and Collusion via Coupling”. In: EC ’23, p. 208. DOI: 10.1145/3580507.3597726. URL: https://doi.org/10.1145/3580507.3597726.

Baradaran, Mehrsa (2019). “Jim Crow Credit”. In: *UC Irvine L. Rev.* 9 (4).

Bardhi, Arjada, Yingni Guo, and Bruno Strulovici (2020). *Early-Career Discrimination: Spiraling or Self-Correcting?* Tech. rep.

Bechavod, Yahav, Chara Podimata, Steven Wu, and Juba Ziani (July 2022). “Information Discrepancy in Strategic Learning”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 1691–1715. URL: https://proceedings.mlr.press/v162/bechavod22a.html.

Becker, Gary S. and Kevin M. Murphy (1988). “A Theory of Rational Addiction”. In: *Journal of Political Economy* 96.4, pp. 675–700. ISSN: 00223808, 1537534X. URL: http://www.jstor.org/stable/1830469 (visited on 01/31/2025).

Benthall, Sebastian, Seda Gürses, and Helen Nissenbaum (Dec. 2017). *Contextual Integrity through the Lens of Computer Science*. Vol. 2. 1. Hanover, MA, USA: Now Publishers Inc., pp. 1–69. DOI: 10.1561/3300000016. URL: https://doi.org/10.1561/3300000016.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2022). “The economics of social data”. In: *The RAND Journal of Economics* 53.2, pp. 263–296. DOI: https://doi.org/10.1111/1756-2171.12407. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1756-2171.12407. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1756-2171.12407.

Bergemann, Dirk, Alessandro Bonatti, Andreas Haupt, and Alex Smolin (2021). “The Optimality of Upgrade Pricing”. In: *Web and Internet Economics: 17th International Conference, WINE 2021, Potsdam, Germany, December 14–17, 2021, Proceedings*. Potsdam, Germany: Springer-Verlag, pp. 41–58. ISBN: 978-3-030-94675-3. DOI: 10.1007/978-3-030-94676-0_3. URL: https://doi.org/10.1007/978-3-030-94676-0_3.

Bergemann, Dirk and Stephen Morris (Mar. 2019). “Information Design: A Unified Perspective”. In: *Journal of Economic Literature* 57.1, pp. 44–95. DOI: 10.1257/jel.20181489. URL: https://www.aeaweb.org/articles?id=10.1257/jel.20181489.

Bergemann, Dirk and Juuso Välimäki (2018). “Bandit Problems”. In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK, pp. 665–670. ISBN: 978-1-349-95189-5. DOI: 10.1057/978-1-349-95189-5_2386. URL: https://doi.org/10.1057/978-1-349-95189-5_2386.

Bernheim, B. Douglas (2016). “The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics”. In: *Journal of Benefit-Cost Analysis* 7.1, pp. 12–68. DOI: 10.1017/bca.2016.5.

Bernheim, B. Douglas and Antonio Rangel (Feb. 2009). “Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics*”. In: *The Quarterly Journal of Economics* 124.1, pp. 51–104. ISSN: 0033-5533. DOI: 10.1162/qjec.2009.124.1.51.

eprint: https://academic.oup.com/qje/article-pdf/124/1/51/5340707/124-1-51.pdf. URL: https://doi.org/10.1162/qjec.2009.124.1.51.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995). “Automobile Prices in Market Equilibrium”. In: *Econometrica* 63.4, pp. 841–890. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2171802 (visited on 01/31/2025).

— (2004). “Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market”. In: *Journal of Political Economy* 112.1, pp. 68–105. ISSN: 00223808, 1537534X. URL: http://www.jstor.org/stable/10.1086/379939 (visited on 01/31/2025).

Berry, Steven T. (1994). “Estimating Discrete-Choice Models of Product Differentiation”. In: *The RAND Journal of Economics* 25.2, pp. 242–262. ISSN: 07416261. URL: http://www.jstor.org/stable/2555829 (visited on 01/31/2025).

Blume, Andreas, Oliver J. Board, and Kohei Kawamura (2007). “Noisy talk”. In: *Theoretical Economics* 2.4, pp. 395–440.

Bogetoft, Peter et al. (2009). “Secure Multiparty Computation Goes Live”. In: *Financial Cryptography and Data Security*. Ed. by Roger Dingledine and Philippe Golle. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 325–343. ISBN: 978-3-642-03549-4.

Bolton, Patrick and Christopher Harris (1999). “Strategic Experimentation”. In: *Econometrica* 67.2, pp. 349–374. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2999588 (visited on 01/31/2025).

Brandt, Felix and Tuomas Sandholm (2005). “Unconditional privacy in social choice”. In: *Proceedings of the 10th Conference on Theoretical Aspects of Rationality and Knowledge*. TARK ’05. Singapore: National University of Singapore, pp. 207–218. ISBN: 9810534124.

— (May 2008). “On the Existence of Unconditionally Privacy-Preserving Auction Protocols”. In: *ACM Trans. Inf. Syst. Secur.* 11.2. ISSN: 1094-9224. DOI: 10.1145/1330332.1330338. URL: https://doi.org/10.1145/1330332.1330338.

Braverman, Mark and Sumegha Garg (2020). *The Role of Randomness and Noise in Strategic Classification*. Tech. rep. arXiv: 2005.08377 [cs.LG]. URL: https://arxiv.org/abs/2005.08377.

Brown, Zach Y. and Alexander MacKay (May 2023). “Competition in Pricing Algorithms”. In: *American Economic Journal: Microeconomics* 15.2, pp. 109–56. DOI: 10.1257/mic.20210158. URL: https://www.aeaweb.org/articles?id=10.1257/mic.20210158.

Budish, Eric, Peter Cramton, Albert Kyle, Jeongmin Lee, and David Malec (Apr. 2023). *Flow Trading*. Tech. rep. National Bureau of Economic Research. DOI: 10.3386/w31098. URL: http://dx.doi.org/10.3386/w31098.

Cai, Yang and Constantinos Daskalakis (2022). “Recommender Systems meet Mechanism Design”. In: EC ’22, pp. 897–914. DOI: 10.1145/3490486.3538354. URL: https://doi.org/10.1145/3490486.3538354.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello (Oct. 2020). “Artificial Intelligence, Algorithmic Pricing, and Collusion”. In: *American Economic Review* 110.10, pp. 3267–97. DOI: 10.1257/aer.20190623. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20190623.

— (2023). *Artificial Intelligence, Algorithmic Recommendations and Competition*. Tech. rep. DOI: 10.2139/ssrn.4448010. URL: http://dx.doi.org/10.2139/ssrn.4448010.

Canetti, Ran, Amos Fiat, and Yannai A. Gonczarowski (2023). *Zero-Knowledge Mechanisms*. Tech. rep. arXiv: 2302.05590 [econ.TH]. URL: https://arxiv.org/abs/2302.05590.

Carroll, Micah, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan (2021). "Estimating and Penalizing Preference Shift in Recommender Systems". In: *Proceedings of the 15th ACM Conference on Recommender Systems*. RecSys '21. Amsterdam, Netherlands: Association for Computing Machinery, pp. 661–667. ISBN: 9781450384582. DOI: 10.1145/3460231.3478849. URL: https://doi.org/10.1145/3460231.3478849.

Cen, Sarah H., Andrew Ilyas, Jennifer Allen, Hannah Li, and Aleksander Madry (2024). "Measuring Strategization in Recommendation: Users Adapt Their Behavior to Shape Future Content". In: EC '24, pp. 203–204. DOI: 10.1145/3670865.3673634. URL: https://doi.org/10.1145/3670865.3673634.

Cen, Sarah H., Andrew Ilyas, and Aleksander Madry (2024). "User Strategization and Trustworthy Algorithms". In: *Proceedings of the 25th ACM Conference on Economics and Computation*. EC '24. New Haven, CT, USA: Association for Computing Machinery, p. 202. ISBN: 9798400707049. DOI: 10.1145/3670865.3673545. URL: https://doi.org/10.1145/3670865.3673545.

Chaney, Allison J. B., Brandon M. Stewart, and Barbara E. Engelhardt (2018). "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys '18. Vancouver, British Columbia, Canada: Association for Computing Machinery, pp. 224–232. ISBN: 9781450359016. DOI: 10.1145/3240323.3240370. URL: https://doi.org/10.1145/3240323.3240370.

Chen, Yiling, Yang Liu, and Chara Podimata (2020). "Learning Strategy-Aware Linear Classifiers". In: 33. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, pp. 15265–15276. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/ae87a54e183c075c494c4d397d126a66-Paper.pdf.

Chen, Yiling, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah (2018). "Strategyproof Linear Regression in High Dimensions". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. EC '18. Ithaca, NY, USA: Association for Computing Machinery, pp. 9–26. ISBN: 9781450358293. DOI: 10.1145/3219166.3219175. URL: https://doi.org/10.1145/3219166.3219175.

Chen, Ying (2011). "Perturbed communication games with honest senders and naive receivers". In: *Journal of Economic Theory* 146.2, pp. 401–424. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2010.08.001. URL: https://www.sciencedirect.com/science/article/pii/S0022053110001134.

Chen, Yiqiu and Alexander Westkamp (2022). "Optimal Sequential Implementation". Working Paper.

Chor, Benny, Mihály Geréb-Graus, and Eyal Kushilevitz (1994). "On the structure of the privacy hierarchy". In: *Journal of Cryptology* 7.1, pp. 53–60.

Chor, Benny and E. Kushilevitz (1989). "A Zero-One Law for Boolean Privacy". In: *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*. STOC '89. Seattle, Washington, USA: Association for Computing Machinery, pp. 62–72. ISBN: 0897913078. DOI: 10.1145/73007.73013. URL: https://doi.org/10.1145/73007.73013.

Cornière, Alexandre de and Greg Taylor (2019). "A model of biased intermediation". In: *The RAND Journal of Economics* 50.4, pp. 854–882. ISSN: 07416261, 17562171. URL: http://www.jstor.org/stable/45219895 (visited on 01/31/2025).

Crawford, Gregory S. and Matthew Shum (2005). "Uncertainty and Learning in Pharmaceutical Demand". In: *Econometrica* 73.4, pp. 1137–1173. DOI: https://doi.org/10.1111/j.1468-0262.2005.00612.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2005.00612.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00612.x.

Cummings, Rachel, Stratis Ioannidis, and Katrina Ligett (July 2015). "Truthful Linear Regression". In: *Proceedings of The 28th Conference on Learning Theory*. Ed. by Peter Grünwald, Elad Hazan, and Satyen Kale. Vol. 40. Proceedings of Machine Learning Research. Paris, France: PMLR, pp. 448–483. URL: https://proceedings.mlr.press/v40/Cummings15.html.

Curmei, Mihaela, Andreas Haupt, Benjamin Recht, and Dylan Hadfield-Menell (2022). "Towards Psychologically-Grounded Dynamic Preference Models". In: *Proceedings of the 16th ACM Conference on Recommender Systems*. RecSys '22. Seattle, WA, USA: Association for Computing Machinery, pp. 35–48. ISBN: 9781450392785. DOI: 10.1145/3523227.3546778. URL: https://doi.org/10.1145/3523227.3546778.

Dai, Jessica, Bailey Flanigan, Nika Haghtalab, Meena Jagadeesan, and Chara Podimata (May 2024). "Can Probabilistic Feedback Drive User Impacts in Online Platforms?" In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, pp. 2512–2520. URL: https://proceedings.mlr.press/v238/dai24b.html.

Damian, Alex, Tengyu Ma, and Jason D. Lee (2021). "Label Noise SGD Provably Prefers Flat Global Minimizers". In: ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. URL: https://openreview.net/forum?id=x2TMPhseWAW.

de Haan, Thomas, Theo Offerman, and Randolph Sloof (2011). "Noisy signaling: Theory and experiment". In: *Games and Economic Behavior* 73.2, pp. 402–428. ISSN: 0899-8256. DOI: https://doi.org/10.1016/j.geb.2011.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0899825611000741.

Deneckere, Raymond and Sergei Severinov (2022). "Signalling, screening and costly misrepresentation". In: *Canadian Journal of Economics/Revue canadienne d'économique* 55.3, pp. 1334–1370. DOI: https://doi.org/10.1111/caje.12614. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/caje.12614. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/caje.12614.

Difallah, Djellel, Elena Filatova, and Panos Ipeirotis (2018). "Demographics and Dynamics of Mechanical Turk Workers". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: Association for Computing Machinery, pp. 135–143. ISBN: 9781450355810. DOI: 10.1145/3159652.3159661. URL: https://doi.org/10.1145/3159652.3159661.

Dong, Jinshuo, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu (2018). "Strategic Classification from Revealed Preferences". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. EC '18. Ithaca, NY, USA: Association for

Computing Machinery, pp. 55–70. ISBN: 9781450358293. DOI: 10.1145/3219166.3219193. URL: https://doi.org/10.1145/3219166.3219193.

Dworczak, Piotr (2020). “Mechanism Design With Aftermarkets: Cutoff Mechanisms”. In: *Econometrica* 88.6, pp. 2629–2661. DOI: https://doi.org/10.3982/ECTA15768. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15768. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15768.

Dwork, Cynthia (2006). “Differential privacy”. In: *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*. ICALP’06. Venice, Italy: Springer-Verlag, pp. 1–12. ISBN: 3540359079. DOI: 10.1007/11787006_1. URL: https://doi.org/10.1007/11787006_1.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006). “Calibrating noise to sensitivity in private data analysis”. In: *Proceedings of the Third Conference on Theory of Cryptography*. TCC’06. New York, NY: Springer-Verlag, pp. 265–284. ISBN: 3540327312. DOI: 10.1007/11681878_14. URL: https://doi.org/10.1007/11681878_14.

Edelman, Benjamin, Michael Ostrovsky, and Michael Schwarz (Mar. 2007). “Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords”. In: *American Economic Review* 97.1, pp. 242–259. DOI: 10.1257/aer.97.1.242. URL: https://www.aeaweb.org/articles?id=10.1257/aer.97.1.242.

Eilat, Ran, Kfir Eliaz, and Xiaosheng Mu (2021). “Bayesian privacy”. In: *Theoretical Economics* 16.4, pp. 1557–1603. DOI: https://doi.org/10.3982/TE4390. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/TE4390. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/TE4390.

Eliaz, Kfir and Ran Spiegler (Sept. 2019). “The Model Selection Curse”. In: *American Economic Review: Insights* 1.2, pp. 127–40. DOI: 10.1257/aeri.20180485. URL: https://www.aeaweb.org/articles?id=10.1257/aeri.20180485.

Eslami, Motahhare, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik (2016). “First I "like" it, then I hide it: Folk Theories of Social Feeds”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: Association for Computing Machinery, pp. 2371–2382. ISBN: 9781450333627. DOI: 10.1145/2858036.2858494. URL: https://doi.org/10.1145/2858036.2858494.

Evaluating Large Language Models Trained on Code (2021). arXiv: 2107.03374 [cs.LG]. URL: https://arxiv.org/abs/2107.03374.

Eysenbach, Benjamin and Sergey Levine (2019). *If MaxEnt RL is the Answer, What is the Question?* Tech. rep. arXiv: 1910.01913 [cs.LG]. URL: https://arxiv.org/abs/1910.01913.

Fan, Lin and Peter W. Glynn (2024). *Diffusion Approximations for Thompson Sampling*. arXiv: 2105.09232 [cs.LG]. URL: https://arxiv.org/abs/2105.09232.

Farber, Henry S. and Robert Gibbons (Nov. 1996). “Learning and Wage Dynamics*”. In: *The Quarterly Journal of Economics* 111.4, pp. 1007–1047. ISSN: 0033-5533. DOI: 10.2307/2946706. eprint: https://academic.oup.com/qje/article-pdf/111/4/1007/5444385/111-4-1007.pdf. URL: https://doi.org/10.2307/2946706.

Fleder, Daniel and Kartik Hosanagar (2009). “Blockbuster Culture’s Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity”. In: *Management Science* 55.5, pp. 697–712. doi: 10.1287/mnsc.1080.0974. eprint: https://doi.org/10.1287/mnsc.1080.0974. url: https://doi.org/10.1287/mnsc.1080.0974.

Franklin, M.K. and M.K. Reiter (1996). "The design and implementation of a secure auction service". In: *IEEE Transactions on Software Engineering* 22.5, pp. 302–312. doi: 10.1109/32.502223.

Frisch, Uriel and Hélène Frisch (1995). "Universality of escape from a half-space for symmetrical random walks". In: *Lévy Flights and Related Topics in Physics*. Ed. by Micheal F. Shlesinger, George M. Zaslavsky, and Uriel Frisch. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 262–268. isbn: 978-3-540-49225-2.

Fukuyama, Francis, Barak Richman, Ashish Goel, Roberta R. Katz, A. Douglas Melamed, and Marietje Schaake (2020). *Middleware for dominant digital platforms: A technological solution to a threat to democracy*. Tech. rep. CyberPolicy Center, Freeman Spogli Institute.

Gemp, Ian, Andreas Haupt, Luke Marris, Siqi Liu, and Georgios Piliouras (2025). *Convex Markov Games: A Framework for Creativity, Imitation, Fairness, and Safety in Multiagent Learning*. arXiv: 2410.16600 [cs.GT]. url: https://arxiv.org/abs/2410.16600.

Ghalme, Ganesh, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld (July 2021). "Strategic Classification in the Dark". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 3672–3681. url: https://proceedings.mlr.press/v139/ghalme21a.html.

Ghosh, Arpita and Aaron Roth (2015). "Selling privacy at auction". In: *Games and Economic Behavior* 91, pp. 334–346. issn: 0899-8256. doi: https://doi.org/10.1016/j.geb.2013.06.013. url: https://www.sciencedirect.com/science/article/pii/S0899825613000961.

Goldfarb, Avi and Verina F. Que (2023). "The Economics of Digital Privacy". In: *Annual Review of Economics* 15.Volume 15, 2023, pp. 267–286. issn: 1941-1391. doi: https://doi.org/10.1146/annurev-economics-082322-014346. url: https://www.annualreviews.org/content/journals/10.1146/annurev-economics-082322-014346.

Goldfarb, Avi and Catherine Tucker (May 2012). "Shifts in Privacy Concerns". In: *American Economic Review* 102.3, pp. 349–53. doi: 10.1257/aer.102.3.349. url: https://www.aeaweb.org/articles?id=10.1257/aer.102.3.349.

Golowich, Louis and Shengwu Li (2022). "On the Computational Properties of Obviously Strategy-Proof Mechanisms". In: arXiv: 2101.05149 [econ.TH]. url: https://arxiv.org/abs/2101.05149.

Gomez-Uribe, Carlos A. and Neil Hunt (Dec. 2016). "The Netflix Recommender System: Algorithms, Business Value, and Innovation". In: *ACM Trans. Manage. Inf. Syst.* 6.4. issn: 2158-656X. doi: 10.1145/2843948. url: https://doi.org/10.1145/2843948.

Goodrow, Cristos (2021). *On YouTube's Recommendation System*. url: https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.

Grigoryan, Aram and Markus Möller (2023). "A Theory of Auditability for Allocation and Social Choice Mechanisms". In: EC '23, p. 815. doi: 10.1145/3580507.3597708. url: https://doi.org/10.1145/3580507.3597708.

Guo, Xiaotong, Andreas Haupt, Hai Wang, Rida Qadri, and Jinhua Zhao (2023). "Understanding multi-homing and switching by platform drivers". In: *Transportation Research Part C: Emerging Technologies* 154, p. 104233. issn: 0968-090X. doi: https://doi.org/10.

1016/j.trc.2023.104233. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2300222X.

Gupta, Aastha (2021). *Incorporating More Feedback Into News Feed Ranking | Meta — about.fb.com*. https://about.fb.com/news/2021/04/incorporating-more-feedback-into-news-feed-ranking/. [Accessed 21-08-2024].

Gupta, Samarth, Gauri Joshi, and Osman Yağan (2020). "Correlated Multi-Armed Bandits with A Latent Random Source". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3572–3576. DOI: 10.1109/ICASSP40776.2020.9054429.

Gutierrez Gallardo, German (2021). *The Welfare Consequences of Regulating Amazon*. DOI: 10.2139/ssrn.3965566. URL: http://dx.doi.org/10.2139/ssrn.3965566.

Hadfield-Menell, Dylan (Aug. 2021). *The Principal-Agent Alignment Problem in Artificial Intelligence*. UCB/EECS-2021-207. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-207.html.

Hakimov, Rustamdjan and Madhav Raghavan (2020). "Transparency in Centralised Allocation". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3572020. URL: http://dx.doi.org/10.2139/ssrn.3572020.

Hall, P. and C.C. Heyde (1980). "3 - The Central Limit Theorem". In: *Martingale Limit Theory and its Application*. Ed. by P. Hall and C.C. Heyde. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, pp. 51–96. DOI: https://doi.org/10.1016/B978-0-12-319350-6.50009-8. URL: https://www.sciencedirect.com/science/article/pii/B9780123193506500098.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters (2016). "Strategic Classification". In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ITCS '16. Cambridge, Massachusetts, USA: Association for Computing Machinery, pp. 111–122. ISBN: 9781450340571. DOI: 10.1145/2840728.2840730. URL: https://doi.org/10.1145/2840728.2840730.

Harstad, Ronald (July 2018). "US Patent No. 10,726,476 B2: Systems and Methods for Advanced Auction Management". 10,726,476 B2.

Hartzell, Olivia and Andreas Haupt (2025). *Platform Preferencing and Price Competition I: Evidence from Amazon*. Working paper. URL: https://hartzell.scholars.harvard.edu/sites/g/files/omnuum5151/files/2025-01/Platform_preferencing_i_statics.pdf.

Haupt, Andreas and Mihaela Curmei (Sept. 2024). *Regret Data Collection Tool*. Version 0.9. DOI: 10.5281/zenodo.13770054. URL: https://doi.org/10.5281/zenodo.13770054.

Haupt, Andreas and Zoë Hitzig (2022). "Contextually Private Mechanisms". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Boulder, CO, USA: Association for Computing Machinery, p. 1144. ISBN: 9781450391504. DOI: 10.1145/3490486.3538259. URL: https://doi.org/10.1145/3490486.3538259.

Haupt, Andreas, Zoë Hitzig, and Jeffrey Gleason (2023). *Response to the Request to Digital Services Act Request for Evidence: Historical Experimental Data*. Submitted to the European Commission. URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3423926%5C%5Fen.

Haupt, Andreas and Aroon Narayanan (2024). "Risk preferences of learning algorithms". In: *Games and Economic Behavior* 148, pp. 415–426. ISSN: 0899-8256. DOI: https://doi.

org/10.1016/j.geb.2024.09.013. URL: https://www.sciencedirect.com/science/article/pii/S089982562400143X.

Haupt, Andreas, Chara Podimata, and Dylan Hadfield-Menell (2023). *Recommending to Strategic Users*. Workshop on the Foundations of Responsible Computing '23. arXiv: 2302.06559 [CS.CY].

Haupt, Andreas Alexander, Nicole Immorlica, and Brendan Lucier (2024). "Certification Design for a Competitive Market". In: *Proceedings of the 25th ACM Conference on Economics and Computation*. EC '24. New Haven, CT, USA: Association for Computing Machinery, p. 852. ISBN: 9798400707049. DOI: 10.1145/3670865.3673593. URL: https://doi.org/10.1145/3670865.3673593.

Hitlin, Paul (2016). *Research in the crowdsourcing age: A case study*. Tech. rep.

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan (2019). "The Disparate Effects of Strategic Manipulation". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 259–268. ISBN: 9781450361255. DOI: 10.1145/3287560.3287597. URL: https://doi.org/10.1145/3287560.3287597.

Ichihashi, Shota (Feb. 2020). "Online Privacy and Information Disclosure by Consumers". In: *American Economic Review* 110.2, pp. 569–95. DOI: 10.1257/aer.20181052. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20181052.

— (2021). "The economics of data externalities". In: *Journal of Economic Theory* 196, p. 105316. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2021.105316. URL: https://www.sciencedirect.com/science/article/pii/S0022053121001332.

Ipeirotis, Panagiotis G. (Mar. 2010). "Demographics of Mechanical Turk". In: CEDER-10-01. URL: https://ssrn.com/abstract=1585030.

Izmalkov, S., S. Micali, and M. Lepinski (2005). "Rational secure computation and ideal mechanism design". In: *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pp. 585–594. DOI: 10.1109/SFCS.2005.64.

Izmalkov, Sergei, Matt Lepinski, and Silvio Micali (2011). "Perfect implementation". In: *Games and Economic Behavior* 71.1. Special Issue In Honor of John Nash, pp. 121–140. ISSN: 0899-8256. DOI: https://doi.org/10.1016/j.geb.2010.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0899825610000758.

Jagadeesan, Meena, Nikhil Garg, and Jacob Steinhardt (2023). "Supply-side equilibria in recommender systems". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc.

Joseph, Matthew, Michael Kearns, Jamie H Morgenstern, and Aaron Roth (2016). "Fairness in Learning: Classic and Contextual Bandits". In: 29. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/eb163727917cbba1eea208541a643e74-Paper.pdf.

Joskow, Paul L., Richard Schmalensee, and Elizabeth M. Bailey (1998). "The Market for Sulfur Dioxide Emissions". In: *The American Economic Review* 88.4, pp. 669–685. ISSN: 00028282. URL: http://www.jstor.org/stable/117000 (visited on 01/31/2025).

Kaelbling, Leslie Pack, Michael L. Littman, and Anthony R. Cassandra (1998). "Planning and acting in partially observable stochastic domains". In: *Artificial Intelligence* 101.1, pp. 99–134. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(98)00023-X. URL: https://www.sciencedirect.com/science/article/pii/S000437029800023X.

Kalvit, Anand and Assaf Zeevi (2021a). *A Closer Look at the Worst-case Behavior of Multi-armed Bandit Algorithms*. Tech. rep. arXiv: 2106.02126 [cs.LG]. URL: https://arxiv.org/abs/2106.02126.

— (2021b). "A closer look at the worst-case behavior of multi-armed bandit algorithms". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc. ISBN: 9781713845393.

Kamenica, Emir (2019). "Bayesian Persuasion and Information Design". In: *Annual Review of Economics* 11.Volume 11, 2019, pp. 249–272. ISSN: 1941-1391. DOI: https://doi.org/10.1146/annurev-economics-080218-025739. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-economics-080218-025739.

Kamenica, Emir and Matthew Gentzkow (Oct. 2011). "Bayesian Persuasion". In: *American Economic Review* 101.6, pp. 2590–2615. DOI: 10.1257/aer.101.6.2590. URL: https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Kartik, Navin (Oct. 2009). "Strategic Communication with Lying Costs". In: *The Review of Economic Studies* 76.4, pp. 1359–1395. ISSN: 0034-6527. DOI: 10.1111/j.1467-937X.2009.00559.x. eprint: https://academic.oup.com/restud/article-pdf/76/4/1359/18358398/76-4-1359.pdf. URL: https://doi.org/10.1111/j.1467-937X.2009.00559.x.

Keller, Godfrey, Sven Rady, and Martin Cripps (2005). "Strategic Experimentation with Exponential Bandits". In: *Econometrica* 73.1, pp. 39–68. DOI: https://doi.org/10.1111/j.1468-0262.2005.00564.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2005.00564.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00564.x.

Kirk, Hannah Rose et al. (2024). "The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models". In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=DFr5hteojx.

Klein, Nicolas and Sven Rady (2011). "Negatively Correlated Bandits". In: *The Review of Economic Studies* 78.2, pp. 693–732. ISSN: 00346527, 1467937X. URL: http://www.jstor.org/stable/23015871 (visited on 01/31/2025).

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan (2024). "The Inversion Problem: Why Algorithms Should Infer Mental State and Not Just Predict Behavior". In: *Perspectives on Psychological Science* 19.5. PMID: 38085919, pp. 827–838. DOI: 10.1177/17456916231212138. eprint: https://doi.org/10.1177/17456916231212138. URL: https://doi.org/10.1177/17456916231212138.

Klug, Daniel, Yiluo Qin, Morgan Evans, and Geoff Kaufman (2021). "Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm". In: *Proceedings of the 13th ACM Web Science Conference 2021*. WebSci '21. Virtual Event, United Kingdom: Association for Computing Machinery, pp. 84–92. ISBN: 9781450383301. DOI: 10.1145/3447535.3462512. URL: https://doi.org/10.1145/3447535.3462512.

Kuckartz, Udo (2019). "Qualitative Text Analysis: A Systematic Approach". In: *Compendium for Early Career Researchers in Mathematics Education*. Ed. by Gabriele Kaiser and Norma Presmeg. Cham: Springer International Publishing, pp. 181–197. ISBN: 978-

3-030-15636-7. DOI: 10.1007/978-3-030-15636-7_8. URL: https://doi.org/10.1007/978-3-030-15636-7_8.

Laibson, David (May 1997). "Golden Eggs and Hyperbolic Discounting*". In: *The Quarterly Journal of Economics* 112.2, pp. 443–478. ISSN: 0033-5533. DOI: 10.1162/003355397555253. eprint: https://academic.oup.com/qje/article-pdf/112/2/443/5291736/112-2-443.pdf. URL: https://doi.org/10.1162/003355397555253.

Landeras, Pedro and J. M. Pérez de Villarreal (2005). "A Noisy Screening Model of Education". In: *LABOUR* 19.1, pp. 35–54. DOI: https://doi.org/10.1111/j.1467-9914.2005.00297.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9914.2005.00297.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9914.2005.00297.x.

Lattimore, Tor and Csaba Szepesvári (July 2020). *Bandit Algorithms*. Cambridge University Press. ISBN: 9781108486828. DOI: 10.1017/9781108571401. URL: http://dx.doi.org/10.1017/9781108571401.

Lee, Angela Y., Hannah Mieczkowski, Nicole B. Ellison, and Jeffrey T. Hancock (Nov. 2022). "The Algorithmic Crystal: Conceptualizing the Self through Algorithmic Personalization on TikTok". In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2. DOI: 10.1145/3555601. URL: https://doi.org/10.1145/3555601.

Lee, Kwok Hao and Leon Musolff (2021). *Entry into two-sided markets shaped by platform-guided search*. Tech. rep.

Levanon, Sagi and Nir Rosenfeld (July 2021). "Strategic Classification Made Practical". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 6243–6253. URL: https://proceedings.mlr.press/v139/levanon21a.html.

Li, Shengwu (Nov. 2017). "Obviously Strategy-Proof Mechanisms". In: *American Economic Review* 107.11, pp. 3257–87. DOI: 10.1257/aer.20160425. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20160425.

Ligett, Katrina and Aaron Roth (2012). "Take it or leave it: running a survey when privacy comes at a cost". In: *Proceedings of the 8th International Conference on Internet and Network Economics*. WINE'12. Liverpool, UK: Springer-Verlag, pp. 378–391. ISBN: 9783642353109. DOI: 10.1007/978-3-642-35311-6_28. URL: https://doi.org/10.1007/978-3-642-35311-6_28.

Lin, Tesary (2022). "Valuing Intrinsic and Instrumental Preferences for Privacy". In: *Marketing Science* 41.4, pp. 663–681. DOI: 10.1287/mksc.2022.1368. eprint: https://doi.org/10.1287/mksc.2022.1368. URL: https://doi.org/10.1287/mksc.2022.1368.

Liu, De and Adib Bagh (2020). "Preserving Bidder Privacy in Assignment Auctions: Design and Measurement". In: *Management Science* 66.7, pp. 3162–3182. DOI: 10.1287/mnsc.2019.3349. eprint: https://doi.org/10.1287/mnsc.2019.3349. URL: https://doi.org/10.1287/mnsc.2019.3349.

Liu, Yang, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes (2017). *Calibrated Fairness in Bandits*. Tech. rep. arXiv: 1707.01875 [cs.LG]. URL: https://arxiv.org/abs/1707.01875.

Mackenzie, Andrew (2020). "A revelation principle for obviously strategy-proof implementation". In: *Games and Economic Behavior* 124, pp. 512–533. ISSN: 0899-8256. DOI:

https://doi.org/10.1016/j.geb.2020.09.010. URL: https://www.sciencedirect.com/science/article/pii/S0899825620301408.

Mackenzie, Andrew and Yu Zhou (2022). “Menu mechanisms”. In: *Journal of Economic Theory* 204, p. 105511. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2022.105511. URL: https://www.sciencedirect.com/science/article/pii/S0022053122001016.

Marlin, Benjamin M. and Richard S. Zemel (2009). “Collaborative prediction and ranking with non-random missing data”. In: *Proceedings of the Third ACM Conference on Recommender Systems*. RecSys ’09. New York, New York, USA: Association for Computing Machinery, pp. 5–12. ISBN: 9781605584355. DOI: 10.1145/1639714.1639717. URL: https://doi.org/10.1145/1639714.1639717.

McKelvey, Richard D. and Thomas R. Palfrey (1995). “Quantal Response Equilibria for Normal Form Games”. In: *Games and Economic Behavior* 10.1, pp. 6–38. ISSN: 0899-8256. DOI: https://doi.org/10.1006/game.1995.1023. URL: https://www.sciencedirect.com/science/article/pii/S0899825685710238.

McMillan, John (Sept. 1994). “Selling Spectrum Rights”. In: *Journal of Economic Perspectives* 8.3, pp. 145–162. DOI: 10.1257/jep.8.3.145. URL: https://www.aeaweb.org/articles?id=10.1257/jep.8.3.145.

Meir, Reshef, Ariel D. Procaccia, and Jeffrey S. Rosenschein (2012). “Algorithms for strategyproof classification”. In: *Artificial Intelligence* 186, pp. 123–156. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2012.03.008. URL: https://www.sciencedirect.com/science/article/pii/S000437021200029X.

Milgrom, Paul and Ilya Segal (2020). “Clock Auctions and Radio Spectrum Reallocation”. In: *Journal of Political Economy* 128.1, pp. 1–31. DOI: 10.1086/704074. eprint: https://doi.org/10.1086/704074. URL: https://doi.org/10.1086/704074.

Milli, Smitha, Luca Belli, and Moritz Hardt (2021). “From Optimizing Engagement to Measuring Value”. In: FAccT ’21, pp. 714–722. DOI: 10.1145/3442188.3445933. URL: https://doi.org/10.1145/3442188.3445933.

Milli, Smitha, John Miller, Anca D. Dragan, and Moritz Hardt (2019). “The Social Cost of Strategic Classification”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, pp. 230–239. ISBN: 9781450361255. DOI: 10.1145/3287560.3287576. URL: https://doi.org/10.1145/3287560.3287576.

Moulin, H. (1980). “On Strategy-Proofness and Single Peakedness”. In: *Public Choice* 35.4, pp. 437–455. ISSN: 00485829, 15737101. URL: http://www.jstor.org/stable/30023824 (visited on 01/31/2025).

Musolff, Leon (2022). “Algorithmic Pricing Facilitates Tacit Collusion: Evidence from E-Commerce”. In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC ’22. Boulder, CO, USA: Association for Computing Machinery, pp. 32–33. ISBN: 9781450391504. DOI: 10.1145/3490486.3538239. URL: https://doi.org/10.1145/3490486.3538239.

Myerson, Roger B (1982). “Optimal coordination mechanisms in generalized principal–agent problems”. In: *Journal of Mathematical Economics* 10.1, pp. 67–81. ISSN: 0304-4068. DOI: https://doi.org/10.1016/0304-4068(82)90006-4. URL: https://www.sciencedirect.com/science/article/pii/0304406882900064.

Nissenbaum, Helen (2004). “Privacy as contextual integrity”. In: *Wash. L. Rev.* 79, p. 119.

Nissim, Kobbi, Claudio Orlandi, and Rann Smorodinsky (2012). "Privacy-aware mechanism design". In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. EC '12. Valencia, Spain: Association for Computing Machinery, pp. 774–789. ISBN: 9781450314152. DOI: 10.1145/2229012.2229073. URL: https://doi.org/10.1145/2229012.2229073.

Nurmi, Hannu and Arto Salomaa (Nov. 1993). "Cryptographic protocols for Vickrey auctions". In: *Group Decision and Negotiation* 2.4, pp. 363–373. ISSN: 1572-9907. DOI: 10.1007/bf01384489. URL: http://dx.doi.org/10.1007/BF01384489.

O'Donoghue, Ted and Matthew Rabin (Mar. 1999). "Doing It Now or Later". In: *American Economic Review* 89.1, pp. 103–124. DOI: 10.1257/aer.89.1.103. URL: https://www.aeaweb.org/articles?id=10.1257/aer.89.1.103.

— (Feb. 2001). "Choice and Procrastination". In: *The Quarterly Journal of Economics* 116.1, pp. 121–160. ISSN: 0033-5533. DOI: 10.1162/003355301556365. eprint: https://academic.oup.com/qje/article-pdf/116/1/121/5461993/116-1-121.pdf. URL: https://doi.org/10.1162/003355301556365.

Ollar, Mariann, Marzena J. Rostek, and Ji Hee Yoon (2017). "Privacy in Markets". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3071374. URL: http://dx.doi.org/10.2139/ssrn.3071374.

Pai, Mallesh M. and Aaron Roth (June 2013). "Privacy and mechanism design". In: *SIGecom Exch.* 12.1, pp. 8–29. DOI: 10.1145/2509013.2509016. URL: https://doi.org/10.1145/2509013.2509016.

Parkes, David C. and Lyle H. Ungar (2000). "Iterative Combinatorial Auctions: Theory and Practice". In: *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press, pp. 74–81. ISBN: 0262511126.

Parkes, David C., Lyle H. Ungar, and Dean P. Foster (1998). "Accounting for Cognitive Costs in On-Line Auction Design". In: *Selected Papers from the First International Workshop on Agent Mediated Electronic Trading on Agent Mediated Electronic Commerce*. AMET '98. Berlin, Heidelberg: Springer-Verlag, pp. 25–40. ISBN: 3540659552.

Patil, Vishakha, Ganesh Ghalme, Vineet Nair, and Y. Narahari (Jan. 2021). "Achieving fairness in the stochastic multi-armed bandit problem". In: *J. Mach. Learn. Res.* 22.1. ISSN: 1532-4435.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger (2017). "On fairness and calibration". In: NIPS'17, pp. 5684–5693.

Poddar, Sriyash, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques (2024). *Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning*. Tech. rep. arXiv: 2408.10075 [cs.LG]. URL: https://arxiv.org/abs/2408.10075.

Pycia, Marek and Madhav Raghavan (2021). *Non-Bossiness and First-Price Auctions*. Tech. rep. DOI: 10.2139/ssrn.3941784. URL: http://dx.doi.org/10.2139/ssrn.3941784.

Pycia, Marek and Peter Troyan (2023). "A Theory of Simplicity in Games and Mechanism Design". In: *Econometrica* 91.4, pp. 1495–1526. DOI: https://doi.org/10.3982/ECTA16310. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16310. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16310.

Pycia, Marek and M. Utku Ünver (2017). “Incentive compatible allocation and exchange of discrete resources”. In: *Theoretical Economics* 12.1, pp. 287–329. DOI: https://doi.org/10.3982/TE2201. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/TE2201. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/TE2201.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn (2023). “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=HPuSIXJaa9.

Rahwan, Iyad et al. (Apr. 2019). “Machine behaviour”. In: *Nature* 568.7753, pp. 477–486. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1138-y. URL: http://dx.doi.org/10.1038/s41586-019-1138-y.

Raman, Narun, Taylor Lundy, Samuel Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz (2024). *STEER: Assessing the Economic Rationality of Large Language Models*. Tech. rep. arXiv: 2402.09552 [cs.CL]. URL: https://arxiv.org/abs/2402.09552.

Roth, Aaron and Grant Schoenebeck (2012). “Conducting truthful surveys, cheaply”. In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. EC ’12. Valencia, Spain: Association for Computing Machinery, pp. 826–843. ISBN: 9781450314152. DOI: 10.1145/2229012.2229076. URL: https://doi.org/10.1145/2229012.2229076.

Roth, Alvin E. and Elliott Peranson (Sept. 1997). “The Effects of the Change in the NRMP Matching Algorithm”. In: *JAMA* 278.9, pp. 729–732. ISSN: 0098-7484. DOI: 10.1001/jama.1997.03550090053032. eprint: https://jamanetwork.com/journals/jama/articlepdf/418088/jama_278_9_032.pdf. URL: https://doi.org/10.1001/jama.1997.03550090053032.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver (May 2004). “Kidney Exchange*”. In: *The Quarterly Journal of Economics* 119.2, pp. 457–488. ISSN: 0033-5533. DOI: 10.1162/0033553041382157. eprint: https://academic.oup.com/qje/article-pdf/119/2/457/5351694/119-2-457.pdf. URL: https://doi.org/10.1162/0033553041382157.

Rothkopf, Michael H., Thomas J. Teisberg, and Edward P. Kahn (1990). “Why Are Vickrey Auctions Rare?” In: *Journal of Political Economy* 98.1, pp. 94–109. DOI: 10.1086/261670. eprint: https://doi.org/10.1086/261670. URL: https://doi.org/10.1086/261670.

Safoury, Laila and Akram Salah (2013). “Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System”. In: *Lecture Notes on Software Engineering*, pp. 303–307. ISSN: 2301-3559. DOI: 10.7763/lnse.2013.v1.66. URL: http://dx.doi.org/10.7763/LNSE.2013.V1.66.

Satterthwaite, Mark A. and Hugo Sonnenschein (Oct. 1981). “Strategy-Proof Allocation Mechanisms at Differentiable Points”. In: *The Review of Economic Studies* 48.4, pp. 587–597. ISSN: 0034-6527. DOI: 10.2307/2297198. eprint: https://academic.oup.com/restud/article-pdf/48/4/587/4523868/48-4-587.pdf. URL: https://doi.org/10.2307/2297198.

Segal, Ilya (2007). “The communication requirements of social choice rules and supporting budget sets”. In: *Journal of Economic Theory* 136.1, pp. 341–378. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2006.09.011. URL: https://www.sciencedirect.com/science/article/pii/S0022053107000154.

Sethuraman, Ramya, Jordi Vallmitjana, and Jon Levin (2019). *Using Surveys to Make News Feed More Personal*. https://about.fb.com/news/2019/05/more-personalized-experiences/. [Accessed 21-08-2024].

Shapley, Lloyd and Herbert Scarf (1974). "On cores and indivisibility". In: *Journal of Mathematical Economics* 1.1, pp. 23–37. ISSN: 0304-4068. DOI: https://doi.org/10.1016/0304-4068(74)90033-0. URL: https://www.sciencedirect.com/science/article/pii/0304406874900330.

Simpson, Ellen, Andrew Hamann, and Bryan Semaan (Jan. 2022). "How to Tame "Your" Algorithm: LGBTQ+ Users' Domestication of TikTok". In: *Proc. ACM Hum.-Comput. Interact.* 6.GROUP. DOI: 10.1145/3492841. URL: https://doi.org/10.1145/3492841.

Siththaranjan, Anand, Cassidy Laidlaw, and Dylan Hadfield-Menell (2024). "Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=0tWTxYYPnW.

Slivkins, Aleksandrs (2019). *Introduction to Multi-Armed Bandits*. Vol. 12. 1–2. Now Publishers, pp. 1–286. DOI: 10.1561/2200000068. URL: http://dx.doi.org/10.1561/2200000068.

Steck, Harald (2018). "Calibrated recommendations". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys '18. Vancouver, British Columbia, Canada: Association for Computing Machinery, pp. 154–162. ISBN: 9781450359016. DOI: 10.1145/3240323.3240372. URL: https://doi.org/10.1145/3240323.3240372.

Sugden, Robert (Sept. 2004). "The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences". In: *American Economic Review* 94.4, pp. 1014–1033. DOI: 10.1257/0002828042002714. URL: https://www.aeaweb.org/articles?id=10.1257/0002828042002714.

— (July 2018). *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford University Press. ISBN: 9780198825142. DOI: 10.1093/oso/9780198825142.001.0001. URL: https://doi.org/10.1093/oso/9780198825142.001.0001.

— (2021). "A response to six comments on The Community of Advantage". In: *Journal of Economic Methodology* 28.4, pp. 419–430. DOI: 10.1080/1350178X.2021.1994634. eprint: https://doi.org/10.1080/1350178X.2021.1994634. URL: https://doi.org/10.1080/1350178X.2021.1994634.

Svensson, Lars-Gunnar (1999). "Strategy-proof allocation of indivisible goods". In: *Social Choice and Welfare* 16.4, pp. 557–567. ISSN: 01761714, 1432217X. URL: http://www.jstor.org/stable/41106323 (visited on 01/31/2025).

Tang, Huan (2019). *The Value of Privacy: Evidence from Online Borrowers*. Tech. rep. Available at https://wpcarey.asu.edu/sites/default/files/2021-11/huan_tang_seminar_paper.pdf.

Weitzman, Martin L. (1979). "Optimal Search for the Best Alternative". In: *Econometrica* 47.3, pp. 641–654. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1910412 (visited on 01/31/2025).

Woodward, Kyle (2020). *Self-Auditable Auctions*. Tech. rep. Available at https://1.618034.com/research/auto/woodward-2020A.pdf.

Zhang, Brian Hu, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen McAleer, Andreas Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm (2024). "Steering No-Regret Learners to a Desired Equilibrium". In: *Proceedings of the 25th ACM Conference on Economics and Computation*. EC '24. New Haven, CT, USA: Association for Computing Machinery, pp. 73–74. ISBN: 9798400707049. DOI: 10.1145/3670865.3673536. URL: https://doi.org/10.1145/3670865.3673536.

Zhang, Brian Hu, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm (2023). "Computing Optimal Equilibria and Mechanisms via Learning in Zero-Sum Extensive-Form Games". In: *Advances in Neural Information Processing Systems 36*.

Zheng, Yong, Mayur Agnani, and Mili Singh (2017). "Identification of Grey Sheep Users by Histogram Intersection in Recommender Systems". In: *Advanced Data Mining and Applications*. Ed. by Gao Cong, Wen-Chih Peng, Wei Emma Zhang, Chengliang Li, and Aixin Sun. Cham: Springer International Publishing, pp. 148–161. ISBN: 978-3-319-69179-4.