Table of Contents

ACM SIGecom Exchanges, Vol. 22, No. 2, March 2025

Editors-in-Chief: Irene Lo and Sam Taggart

Communications Team: Yang Cai, Kira Goldner, and Jinzhao Wu

ACM Staff: Irene Frawley

Editors' Introduction

IRENE LO
Stanford University
and
SAM TAGGART
Oberlin College

With spring around the corner, it is time for another issue of the SIGecom exchanges. As is tradition, we start the winter issue with the 2025 job candidate profiles. This is the ten-year anniversary of the practice; thanks to Vasilis Gkatzelis and Jason Hartline for compiling this useful resource each year. The rest of the issue contains lots of great research-related content: three research letters, three surveys, and an annotated reading list. We preview each below.

In the first of the letters, Tesary Lin and Avner Strulov-Shlain summarize their paper "Choice Architecture, Privacy Valuations, and Selection Bias in Consumer Data," which won best paper in the Empirics track at EC 2023, and has subsequently been accepted to *Marketing Science*. The paper addresses a ubiquitous concern for data collection in online platforms: when users share data by choice, this introduces selection bias in the gathered data. Through an experiment, they study how framing and pricing the choice to share data can impact the volume and biasedness of the resulting dataset.

The next letter, from Laura Doval and Alex Smolin, summarizes recent work in the *Journal of Political Economy* on the welfare impact of information design. They consider settings such as market segmentation where information disclosed about a population can impact population welfare e.g. through downstream decisions. Different disclosure policies, modeled in this work in terms of Bayesian persuasion, lead to different profiles of welfare across individuals in the population. The paper characterizes and studies the set of implementable welfare profiles.

A position piece from Soheil Feizi, Mohammadtaghi Tajiaghayi, Keivan Rezaei, and Suho Shin rounds out the letters in this issue. They consider the problem of integrating online advertising systems into information-seeking tools based on Large Language Models (LLMs). They argue that existing frameworks for online advertising systems in sponsored search are not directly applicable to LLM advertisement, and propose a novel framework to address the unique challenges that arise. The piece raises intriguing questions about design decisions and technical challenges that may be of interest to both academics and practitioners.

The first of our three surveys, by Chara Podimata, overviews the literature on strategic classification. This problem assumes the data points in a classification problem are self-interested agents who can modify their features to change their label. This leads to novel questions of interest to learning theorists, mechanism designers, and algorithmic fairness researchers alike. The survey gives a compre-

---

Author's address: ilo@stanford.edu, staggart@oberlin.edu.

hensive and well-organized overview of the large volume of recent work on the topic.

Next is a survey on automated mechanism design, contributed by Michael J. Curry, Zhou Fan, Yanchen Jiang, Sai Srivatsa Ravindranath, Tonghan Wang, and David Parkes. Their survey covers recent progress using optimization techniques — especially from deep learning — to design mechanisms that are nearly revenue-optimal and exactly or approximately strategyproof. The mechanisms obtained can be useful out of the box to practitioners, and can additionally be a source of structural insights for a theorist studying multiparameter mechanism design.

A survey on Randomized Apportionment by Haris Aziz concludes our collection of surveys. Apportionment is the problem of allocating seats to political parties or representatives to states in proportion to their size. Unlike traditional deterministic methods, randomized apportionment is able to simultaneously achieve multiple desirable fairness criteria, such as exactly proportional representation of groups ex ante and almost-proportional representation ex post. The survey gives an overview of this growing field by highlighting the breadth of methods and axiomatic properties that are enabled by introducing randomization.

Finally, this issue also includes an annotated reading list by Georgios Papasotiropoulos and Ulrike Schmidt-Kraepelin on liquid democracy. Liquid democracy is a hybrid approach to voting that allows voters to either cast their own votes or delegate their voting power to trusted individuals. The list gives an overview of recent work that reflects the breadth in models and methodologies being used in this active research area.

This issue marks a transition between information directors for the SIG. We'd like to thank outgoing director Yannai Gonczarowski one last time for his exceptional service in the role. We also welcome the incoming communications team taking his place, including communications chair Yang Cai, technical lead Jinzhao Wu, and social media chair Kira Goldner. Their help publishing this issue is greatly appreciated. As always, please continue to volunteer letters, surveys, annotated reading lists or position papers. We hope you enjoy this issue.

SIGecom Job Market Candidate Profiles 2025

This is the ninth annual collection of profiles of the junior faculty job market candidates of the SIGecom community. The forty candidates for 2025 are listed alphabetically and indexed by research areas that define the interests of the community. The candidates can be contacted individually, or collectively via the moderated mailing list ecom-candidates2025@acm.org.

–Vasilis Gkatzelis and Jason Hartline

Fig. 1. Generated using the research summaries of the candidates.

Contents

JERRY ANUNROJWONG (Homepage, CV, Google Scholar)

Thesis: Essays in Robust Auction Design and Electricity Markets ('25)

Advisor: Omar Besbes and Santiago R. Balseiro, Columbia University

Brief Biography: Jerry Anunrojwong is a Ph.D. candidate in the Decision, Risk, and Operations division at Columbia Business School, co-advised by Omar Besbes and Santiago R. Balseiro. He has been recognized with several honors, including being a finalist in the 2022 George Nicholson Student Paper Competition.

Research Summary: I am interested in *market design*, broadly construed. A common thread throughout my research is the critical role of participants' strategic behavior and their interplay with constraints on market structure and information access. In particular, my research is motivated by three themes in markets: (1) *robustness*, (2) *information design*, and (3) applications to *energy and sustainability*.

Energy. In [1], I study the cost of centralized versus decentralized batteries in energy markets. Privately-owned batteries aim to maximize profit, which may not align with the system efficiency goal of minimizing cost. Our model identifies three ways that batteries can distort the market. We also quantify the cost of distortions and calibrate our model with real data. In California (resp. Texas), we have a 15% (resp. 25%) difference in possible cost reductions. These strategic costs could be significant, but they pale in comparison to the cost reduction achieved by having enough batteries in the system. We also show that certain regulations can backfire, so market power mitigation protocols needs to be carefully designed.

Robustness. In [2], I consider a seller optimizing over randomized DSIC mechanisms to minimize the worst-case gap (regret or ratio) between mechanism revenue and the benchmark, where the only thing known about the value distribution of n buyers is that it is i.i.d. and its support is on $[a, b]$. I show that if a/b is below a threshold, second-price auctions (SPA) is optimal; if a/b is above another threshold, a new class of mechanisms I call pooling auctions (POOL) is optimal; if a/b is between the two thresholds, a randomization between SPA and POOL is optimal.

Information Design. In [3], I study the effectiveness of information design in reducing congestion in social services catering to users with varied levels of need. Each arriving user decides either to wait for the service by joining an unobservable FCFS queue, or to leave and get her outside option. I show that with enough heterogeneity in need, information design not only Pareto dominates full-info and no-info mechanisms, in some regimes it achieves the same welfare as the "first-best."

Representative Papers:

[1] Battery Operations in Electricity Markets: Strategic Behavior and Distortions (Under Review at *Management Science*), with S. Balseiro, O. Besbes, B. Xu.

[2] Robust Auction Design with Support Information (Minor Revision at *Management Science* & EC 2023), with S. Balseiro, O. Besbes.

[3] Information Design for Congested Social Services: Optimal Need-Based Persuasion (*Management Science* 2022 & EC 2020), with K. Iyer, V. Manshadi.

ESHWAR RAM ARUNACHALESWARAN (Homepage, CV, Google Scholar)

Thesis: Learning, Games and Fairness: Algorithms for Decision-Making in Complex Environments ('25)

Advisors: Sampath Kannan, Anindya De, University of Pennsylvania

Brief Biography: I'm a final year PhD student at the University of Pennsylvania, working on problems in the intersection of algorithmic game theory and online learning. In particular, I study (learning) algorithms as strategies for repeated games and the dynamics they induce. My research has been supported by a gift from AWS AI for research in Trustworthy AI. Currently, I'm visiting the Simons Institute, UC Berkeley, participating in the year long program on transformers and Language models, where I seek to find problems in the intersection of transformers and game theory.

Research Summary: My research focuses on the interactions between learning algorithms in strategic settings. I aim to understand the strategic responses elicited by these algorithms and the resulting dynamics. Central to my inquiry is the question: what are good learning algorithms for strategic interactions? My work entails evaluating learning algorithms based on how they shape time-dynamic, adaptive best-responses from the other player under varying payoff structures. The questions I address in my work have immense theoretical significance for understanding the field of online learning algorithms and how it intersects with dynamics in games.

Key contributions include defining and analyzing the notion of Pareto-optimality in learning settings [1], optimal commitment in repeated games against a single opponent [2]/ distributions of opponents (work in submission), studying the phenomenon of algorithmic collusion in competitive markets [4], and developing novel insights into the non-manipulability of learning algorithms in repeated games (ongoing work). Through these works, I have demonstrated the unique properties and strategic implications of various learning algorithms, such as no-swap-regret algorithms, and provided new tools for analyzing learning dynamics in games.

A central theme of my research is the use of games as a lens to study learning algorithms, often revealing deeper properties of the algorithms themselves. For example, our results highlight a fundamental distinction between FTRL and no-swap-regret algorithms, driven by the way FTRL algorithms can be exploited in repeated games due to their inherent memory. We also prove the asymptotic equivalence of all no-swap-regret algorithms in strategic environments, a particularly significant result in the light of recent breakthrough results about NSR algorithms.

Representative Papers:

[1] Pareto-Optimal Algorithms for Learning in Games (EC 24)
with Natalie Collina and Jon Schneider

[2] Efficient Stackelberg Strategies for Finitely Repeated Games (AAMAS 23)
with Natalie Collina and Michael Kearns

[3] Oracle Efficient Algorithms for Groupwise Regret (ICLR 24)
with Krishna Acharya, Sampath Kannan, Aaron Roth and Juba Ziani

[4] Algorithmic Collusion Without Threats (ITCS 25)
with Natalie Collina, Sampath Kannan, Aaron Roth, Juba Ziani

SANJAY CHANDLEKAR (Homepage, CV, Google Scholar)

Thesis: Autonomous Broker for Smart Grids: Auction Theory, Broker Design and Bidding Strategies (2024)

Advisor: Sujit Gujar, IIIT Hyderabad

Brief Biography: Sanjay is a final year PhD student in the Machine Learning Lab at IIIT Hyderabad, where he is advised by Professor Sujit Gujar. He has been working with TCS Research Labs since 2018 in the Data and Decision Science group. His research interests revolve around Game Theory, Auction Theory, Reinforcement Learning and Deep Learning. Sanjay holds a bachelor's degree in Computer Science from Nirma University and is a recipient of the Gold Medal of the program.

Research Summary: My research focus has been on optimizing bidding strategies in periodic double auctions (PDAs), specifically in the context of smart grids, aiming to improve smart grids' economic and functional efficiency from the distribution company's perspective. Smart grids operate across three key markets: wholesale, tariff, and balancing. Brokers must procure electricity via day-ahead PDAs from the wholesale market and sell it to customers through competitive tariff contracts that incentivize reduced energy consumption during peak times, thereby mitigating peak demands. Towards this, my work addresses challenges like minimizing procurement costs to buy electricity from the PDAs (most relevant for SIGecom), designing optimal tariff contracts to build customer portfolios to sell procured electricity and mitigating the recurring problem of peak demands in smart grids.

To minimize procurement costs, we design bidding strategies for day-ahead PDAs to help brokers procure electricity at the most economical prices. We leverage auction theory and game theory to analyze agent equilibrium behaviour in double auctions, including single-item, multi-item, single-shot[1, 3], and sequential auction[2, 3] settings. However, theoretical analysis either becomes intractable with the increase in the number of players or requires complete information assumption. To overcome the theoretical limitations and make the bidding strategies applicable to real-world scenarios, we implement these strategies in broker agents using reinforcement learning techniques[1, 3, 4]. We show that the learning-based strategies can effectively converge to theoretical equilibria and can be easily extended for more complex real-world scenarios. With the help of these bidding strategies along with the tariff module, our broker agent, VidyutVanika, became the winner of the International smart grid competition (PowerTAC) in 2021 and 2022.

Representative Papers:

[1] Multi-unit Double Auctions: Equilibrium Analysis and Bidding Strategy using DDPG in Smart-grids (AAMAS'22) with E. Subramanian, S. Bhat, P. Paruchuri, and S. Gujar

[2] Optimizing Prosumer Policies in Periodic Double Auctions Inspired by Equilibrium Analysis (IJCAI'24) with B. Manvi, and E. Subramanian

[3] Equilibrium Analysis and Strategic Bidding for Buyers in Multi-unit PDAs (Under Review in AIJ'24) with B. Manvi, E. Subramanian, and S. Gujar

[4] A Novel Bidding Strategy for PDAs using MCTS in Continuous Action Spaces (EMAS'24) with E. Subramanian

THÉO DELEMAZURE (Homepage, CV, Google Scholar)

Thesis: Small Change in Expressiveness, Big Change in Outcome Quality; Analysing Voting with Axioms and Data. ('25)

Advisor: Jérôme Lang and Dominik Peters (Paris Dauphine University)

Brief Biography: I am a fourth-year Ph.D. candidate at Paris Dauphine University - PSL. I hold a Master's degree in AI and a Diploma from école Normale Supérieure (ENS). During my studies, I completed internships at NYU, Nokia Bell Labs, and TU Berlin. My research focuses on voting theory and, more broadly, on social choice theory. I actively promote new ideas from the COMSOC community online, particularly related to electoral reforms.

Research Summary: My research focuses on voting theory. In one of my most recent works [1], we considered generalizations of the Instant Runoff Voting (IRV) rule to weak order preferences (i.e., people might have indifferences in their rankings). We defined the Approval-IRV rule, which generalizes IRV, and showed with axiomatic characterizations that this generalization is the only one that satisfies interesting normative properties. We also studied how this rule would behave using synthetic and real data.

Most of my works [1,2,3,5] follow the same structure: given a model with preferences (e.g., weak orders), and a question (e.g., how to generalize IRV?), what are the different solutions and how do we differentiate between them? The last part generally involves an *axiomatic analysis* of the solutions (often with impossibility and characterization results) and always an *experimental analysis*, in which we apply the different solutions to synthetic and real datasets.

To collect more datasets on which we can test the proposed solutions, I like to conduct experimental surveys in which we ask participants how they would have voted in specific elections (for instance, a presidential one) with alternative voting methods, such as approval voting or IRV. Another important goal of these surveys is to make participants more aware of the alternatives that exist for the voting systems we use. I was part of such a survey during the 2022 French presidential election and led one in 2024 for the European election in France. We designed a website from scratch for each survey and gathered several thousand responses. This also helped us gather interesting feedback from the participants about the different voting methods we suggested. These surveys are essential to my research, as I use at least one dataset collected from them in all of my projects.

Representative Papers:

[1] Generalizing Instant Runoff Voting to Allow Indifferences (EC 2024) with D. Peters

[2] Comparing Ways of Obtaining Candidate Orderings from Approval Ballots (IJCAI 2024) with C. Dong, D. Peters, M. Tydrichova

[3] Selecting the Most Conflicting Pair of Candidates (IJCAI 2024) with Ł. Janeczko, A. Kaczmarczyk and S. Szufa

[4] Independence of Irrelevant Alternatives under the Lens of Pairwise Distortions (AAAI 2024) with J. Lang and G. Pierczynski

[5] Approval With Runoff (IJCAI 2022) with J. Lang, J.-F. Laslier, R. Sanver

PETER DOE (Homepage, CV)

Thesis: Two-Sided Matching with Constraints: Theory-Driven Solutions ('25)

Advisors: Luciano Pomatto, Caltech; Federico Echenique, UC Berkeley

Brief Biography: I am a Ph.D. candidate in Social Sciences at the California Institute of Technology. I earned my B.B.A. from Baylor University in 2020 with majors in Mathematics, Economics, and Statistics. I have variety of interests at the intersection of computer science and economics, including matching, social choice, and algorithmic game theory.

Research Summary: I am a microeconomic theorist working on algorithmic market design, in particular two-sided matching. My research revolves around broadening classic matching models to incorporate matching activity outside of the regular market. I provide actionable market interventions grounded in theory.

In my job market paper [1], I propose a new solution concept for matching markets when some agents have already found partners. This initial match endows market participants with rights: every participant has the right to remain with her initial partner. Tension arises because an agent may wish to abandon his initial partner, but that requires his initial partner's approval. I propose a new equilibrium solution requiring that an agent can only object to a match if her initial partner agrees to the objection. I then present an algorithm that constructs such a match. My algorithm generalizes the Deferred Acceptance and Top Trading Cycles algorithms. I show my algorithm is immune to a variety of misreporting strategies, and it has applications to numerous markets including the resident-to-hospital match, college admissions, school choice, and labor markets.

In a second paper [2] I investigate the distributional impact of agents preempting a centralized marketplace. The motivation behind this the empirical observation that many matching markets suffer from temporal unraveling, which is when matching agreements are made earlier and earlier in time. I show that less-desirable agents benefit from making earlier offers. I present a two-period model with two hospitals and a continuum of doctors. In the first period, doctors are uncertain about their preferences over hospitals, and hospitals can make exploding offers to the doctors. In the second period, doctors learn their preferences and a centralized clearinghouse coordinates the match between the hospitals and the (remaining) doctors. I show that the ex ante less popular hospital drives both hospitals to make exploding offers in the first period, which results in a match that is inefficient ex post. My result explains why less desirable programs are perceived as being harmed by the recent implementation of an "All-In" policy within the NRMP.

Representative Papers:

[1] Matching with a Status Quo: The Agreeable Core (poster at EC, 2024; working paper)

[2] Ranked-to-Match: The Effects of Early Matching in the NRMP (working paper)

SOROUSH EBADIAN (Homepage, CV, Google Scholar)

Thesis: Designing Fair and Socially-Aligned Decision-Making Mechanisms ('25)

Advisor: Nisarg Shah, University of Toronto

Brief Biography: Soroush is a PhD candidate in the Department of Computer Science at the University of Toronto. In fall 2023, he was a visitor at Harvard University, hosted by Ariel D. Procaccia. He holds an Ontario Graduate Fellowship (2023-24) and multiple departmental awards. Previously, he obtained a B.Sc. in Computer Engineering from Sharif University of Technology in 2020 and was a gold medalist in the Iranian National Olympiad in Informatics (2014).

Research Summary: My research lies at the intersection of computer science and economics, focusing on designing collective decision-making systems that are aligned with societal goals, ensuring fairness and economic efficiency.

Democratic systems. How can we best aggregate ranked preferences to select from a set of candidates? While ranked preferences often fail to capture voters' preference intensities, my work [1] resolves the long-standing open problem of identifying the optimal voting rule with minimum loss in welfare. I have also explored various elicitation formats and the simplicity and explainability of voting rules.

Sortition, an ancient democratic process, has been revived to select citizens' assemblies worldwide, often using uniform random selection. While this method ensures fairness, does it truly represent the population? In [2], we propose a new measure of representation, identifying cases where uniform random selection is fair and representative, and in others, designing algorithms that improve the trade-off.

Resource and task allocation. How can we fairly and efficiently divide goods or chores among people? For goods, the state-of-the-art method guarantees approximate envy-freeness and Pareto efficiency. For chores, my work [3] makes the first non-trivial progress to this problem when individuals classify tasks as easy or difficult. Our algorithm, conjectured to work for all instances, is now deployed to the not-for-profit website Spliddit.org. Moreover, in [4], we propose a new fairness notion, inspired by envy-freeness, that applies to a broad range of collective decision-making settings from voting to participatory budgeting and peer review.

Pluralistic AI alignment. My research also addresses aligning AI agents with multiple individuals having diverse preferences and goals [5]. By using insights from social choice and Markov decision processes, we develop methods for aggregating individual policies into a desirable collective policy.

Representative Papers:

[1] Optimized Distortion and Proportional Fairness in Voting (EC'22 and ACM TEAC) with A. Kahng, D. Peters, and N. Shah

[2] Is Sortition Both Representative and Fair? (NeurIPS'22) with G. Kehne, E. Micha, A. D. Procaccia, and N. Shah

[3] How to Fairly Allocate Easy and Difficult Chores (AAMAS'22) with D. Peters and N. Shah

[4] Harm Ratio: A Novel and Versatile Fairness Criterion (EAAMO'24) with R. Freeman and N. Shah

[5] Policy Aggregation (NeurIPS'24) with P. A. Alamdari and A. D. Procaccia

FRANCESCO FABBRI (Homepage, CV)

Thesis: Essays on Attention in Economics ('25)

Advisor: Pietro Ortoleva, Princeton University

Brief Biography: I am a PhD candidate in Economics at Princeton University. My research investigates theoretical problems related to information and uncertainty.

Research Summary: I am interested in addressing two broad questions: (*i*) the effect that information—whether exogenous or endogenously acquired— has on individuals' beliefs and subsequent choices; (*ii*) the design of decision-making models that rationalize behavioral anomalies, observed in the lab or theorized in existing models. I relate these questions to my papers, which I organize into three categories.

Rational Inattention. This framework models agents facing the trade-off between processing more information to improve decisions and saving on information costs. In [1], a producer sets the quality of its product before offering it for a fixed price to a consumer, who processes information about quality, incurring entropy costs. As the consumer becomes more attentive, quality rises, but trade frequency falls as high quality is produced less often. When attention costs vanish, only low quality is provided, causing market failure. [2] analyzes competitive firms' pricing strategies when consumers learn about prices only by paying attention to them. Two forces are at play: inattention may cause market failure, as firms charge high prices, but competition helps keep prices in check. When attention costs are sufficiently high, the effect of competition dominates, increasing trade and benefiting industries that obtain higher profits by competing rather than colluding.

Dynamic Games. The notion of Nash equilibrium is ubiquitous in the game theoretic analysis. However, establishing equilibrium existence has proven elusive in general settings, which are affected by dynamic considerations or where uncertainty entails uncountably infinite states. In [3], we relate equilibrium existence to players making imprecise observations about the history of the game and show that equilibrium exists in general games whenever any amount of idiosyncratic noise is included in players' observations.

Ambiguity aversion. This literature aims to accommodate the Ellsberg paradox and the related experimental evidence, which falsifies expected utility theory. In [4], I characterize the behavior of inattentive and ambiguity averse agents, showing it is consistent with violations of invariance under compression, a property that has been experimentally challenged. [5] axiomatically characterizes preferences that display less aversion to ambiguity as welfare improves.

Representative Papers:

[1] Attention Holdup (Job Market Paper)

[2] Competing to Commit: Markets with Rational Inattention (American Economic Review, 114, no.1, (2024):285-306), with C. Cusumano and F. Pieroth

[3] Stochastic Games with Noisy Informational Asymmetries (EC '24, Extended Abstract), with S. Moroni

[4] Rational Inattention with Ambiguity Aversion (Working Paper)

[5] Absolute and Relative Ambiguity Attitudes (Working Paper), with G. Principi and L. Stanca

ALIREZA FALLAH (Homepage, CV, Google Scholar)

Thesis: Algorithmic Interactions with Strategic Users: Incentives, Interplay, and Impact

Advisor: Asuman Ozdaglar, MIT

Brief Biography: Alireza Fallah is a postdoctoral researcher at UC Berkeley, hosted by Michael Jordan. In the summer of 2023, he obtained his Ph.D. in Electrical Engineering and Computer Science from MIT, where he worked with Asu Ozdaglar and Daron Acemoglu. He spent the fall of 2023 as the Gamelin Postdoctoral Fellow at the Simons Laufer Mathematical Sciences Institute (formerly MSRI), where he was a member of the Mathematics and Computer Science of Market and Mechanism Design program. He has received a number of awards and fellowships, including the honorable mention at the ACM SIGecom Doctoral Dissertation Award, the Ernst A. Guillemin MIT M.Sc. Thesis Award, the Apple Scholars in AI/ML Ph.D. Fellowship, the MathWorks Engineering Fellowship, and the Siebel Scholarship.

Research Summary: My research bridges economics, machine learning theory, and optimization to tackle the challenges arising from the interaction between machine learning (ML) algorithms and human behavior. Below are two key research directions I have explored:

• Data, the fuel that powers ML models and algorithms, is typically collected from users, which has raised various concerns, ranging from privacy to social welfare. My research addresses several key challenges that arise in this context, including: (i) understanding the impact of the emergence of data marketplaces and data monetization on user welfare and how regulations can improve it [1]; and (ii) designing mechanisms that provide both compensation and privacy guarantees to mitigate the issue of free-riding and incentivize data sharing [2].

• There is growing demand to embed societal values like privacy, fairness, and safety into ML models. While much research focuses on improving algorithms, the interaction between these models and the people they impact, particularly when individuals adjust their behavior in response to regulations, is underexplored. My research investigates how human incentives intersect with regulatory interventions, such as safety inspections in contract design [3] or fairness constraints in dynamic auction design [4], identifying when these interventions benefit users and when strategic behavioral shifts may cause unintended consequences.

Representative Papers:

[1] On Three-Layer Data Markets (Under review, shorter version accepted for oral presentation at the ICML Workshop on Agentic Markets), with A. Makhdoumi, A. Malekian, and M. I. Jordan

[2] Optimal and Differentially Private Data Acquisition: Central and Local Mechanisms (Operations Research 2023 & EC 2022), with A. Makhdoumi, A. Malekian, and A. Ozdaglar

[3] Contract Design with Safety Inspections (EC 2024), with M. I. Jordan.

[4] Fair Allocation in Dynamic Mechanism Design (NeurIPS 2024), with A. Ulichney and M. I. Jordan

KARL FEHRS (Homepage, CV)

Thesis: Optimization, Learning, and Fairness in Voting ('25)

Advisor: Ioannis Caragiannis, Aarhus University

Brief Biography: I am a Ph.D. student at the CS department of Aarhus University, Denmark, graduating in spring 2025. My research focus has been on topics at the interface of computer science and economics, usually motivated by questions in social choice theory. During my Ph.D. studies, I had the great honour to be hosted by Prof. Ariel Procaccia at Harvard University for a six month research stay. I hold a B.Sc. degree in CS from Goethe University Frankfurt, Germany, where I also took up my Master's studies in CS. I obtained my M.Sc. degree in CS from Aarhus University as part of my Ph.D. program. I also possess a German diploma in Business Law and work experience in the financial industry.

Research Summary: My research is primarily aimed at understanding decision making processes which are typically framed as voting problems. Specifically, this lead me to study approval-based committee scoring rules [1], metric multiwinner voting [2], and single winner voting under random utilities [3].

On the one hand, my work followed the well-established approach of viewing decision making processes as optimization problems which has motivated the notion of *distortion* in voting. Beyond the classic distortion setting, my research was directed at understanding the additional power of allowing a limited number of queries to the exact cardinal preferences of the agents [2,3]. We introduced a novel average-case notion of distortion in a random utility model [3]. In this model, we were able to show that as few as one query per agent can drastically improve the average distortion to an extend that is not achievable under the traditional worst-case notion of distortion.

In addition, my research employed established CS techniques from learning theory (PAC-learning [1]), complexity theory (parameterized complexity [1]), and the study of algorithms (clustering [2]). Most recently, I have been working in the sortition model of voting theory where my efforts are focused on leveraging insights from the study of clustering algorithms. I expect to make a contribution to the growing literature on the interplay of these two models (sortition/clustering) soon.

In my future research, I would like to further investigate distortion or similar measures of optimality under randomized preferences (e.g., utilities drawn from probability distributions, Mallow's rankings). Exploring this direction outside of the voting setting, e.g., for matching markets, interests me as well. I would also be happy to pursue other directions in computational social choice, e.g., the study of axiomatic properties such as proportionality in voting and beyond.

Representative Papers:

[1] The Complexity of Learning Approval-Based Multiwinner Voting Rules (AAAI-22) with I. Caragiannis

[2] Low-Distortion Clustering with Ordinal and Limited Cardinal Information (AAAI-24) with J. Burkhardt, I. Caragiannis, M. Russo, C. Schwiegelshohn, and S. Shyam

[3] Beyond the Worst Case: Distortion in Impartial Culture Electorates (WINE-24) with I. Caragiannis

SIMON FINSTER (Homepage, CV)

Thesis: Essays on Competitive and Strategic Bidding in Multi-Object Auctions ('22)

Advisor: Paul Klemperer, University of Oxford

Brief Biography: I'm a postdoctoral fellow at CREST (Paris) and the Inria group FairPlay, hosted by Patrick Loiseau and Bary Pradelski. Previously, I was an Associate Fellow at the Simons Laufer Mathematical Institute in Berkeley, CA. I completed my PhD in Economics at the University of Oxford, Nuffield College, a Masters at the Paris School of Economics (APE), and an undergraduate degree in Industrial Engineering at the KIT (Germany).

Research Summary: My main research agenda focuses on equity and fairness concerns in market design.

In my job market paper [1], we initiate the study of surplus equity in auctions for multiple items. We characterize the surplus-equitable mechanism and develop prior-free results on equity-preferred mixed pricing. Our results imply simple policy recommendations for electricity markets, e.g., the equity benefits of a tax discussed by the New Zealand Electricity Authority (2014), and for auctions of carbon emission permits, treasury bonds, and beyond.

I have also proposed and studied a framework for indivisible goods markets with preferences over how goods or services should be distributed among buyers [2]. In related work, we have explored the theoretical foundations of markets with budget-constrained buyers, with applications to the potentially exploitative behavior of digital monopolies [3].

I enjoy bringing market design to the field and the lab, aiming to create productive channels between theoretical and applied work. As such, we have devised pooled testing mechanisms for infectious diseases, in populations with heterogeneous social welfare weights [4]. I am also conducting a large-scale virtual lab experiment (>1100 participants) that sheds light on bidding behavior in auctions for substitutes. In my postdoc research group and beyond, I collaborate with computer scientists and mathematicians. In current work, we advance the understanding of strategic behavior in complex auctions using methods from machine learning.

Representative Papers:

[1] Equitable Pricing in Auctions (Working Paper, 2024). *Job Market Paper.*
with P. Loiseau, S. Mauras, M. Molina, and B. Pradelski.

[2] Selling Multiple Complements with Packaging Costs. (Working Paper, 2024). *Young Economist Essay Award Finalist EARIE 2021.*

[3] Substitutes Markets with Budget Constraints: Solving for Competitive and Optimal Prices (WINE 2023). *R&R at Theoretical Economics.*
with P. W. Goldberg, and E. Lock.

[4] Welfare-Maximizing Pooled Testing (EC 2023). *Exemplary paper award in applied modeling track.*
with M. González Amador, E. Lock, F. Marmolejo-Cossío, E. Micha, and A. Procaccia.

MATTHIAS GREGER (Homepage, CV, Google Scholar)

Thesis: Collective Choice from the Probability Simplex with Application to Donor Coordination ('25)

Advisor: Felix Brandt, Technical University of Munich (TUM)

Brief Biography: I am a fourth-year PhD student at TUM. My research interests comprise various topics from social choice and game theory with a focus on *how to reach and justify fair outcomes*. Together with my co-authors, I received the Best Student Paper Award at WINE 2021. I hold a B.Sc. and M.Sc. in Mathematics from TUM. During my studies, I was supported by the German Academic Scholarship Foundation. At the moment, I am in the process of finishing my dissertation.

Research Summary: My research concerns the aggregation of preferences over a convex set of outcomes, e.g., the probability simplex ([2]), and is applicable but not limited to donor coordination. Specifically, I aim at (i) finding and computing fair and Pareto optimal outcomes, (ii) explaining decisions and singling out the contribution of each individual, and (iii) investigating the stability of outcomes. To achieve these goals, I apply methods from game theory, optimization, and dynamical systems.

Donor coordination deals with the problem of distributing donations from a set of agents among a set of public projects/goods. On the one hand, a central coordination mechanism should distribute the total budget in a Pareto optimal and fair way (or at least give such recommendations). On the other hand, agents want to choose the distributions of their individual donations depending on the distributions of others and the "nature" of the projects. In detail, projects might be interpreted as *substitutes* ([4]) or *complements* ([1,3]). For both models, we prove that maximizing a weighted product of the agents' utilities leads to Pareto optimal and fair outcomes.

In [4], we show that this mechanism incentivizes agents to contribute to public goods even when private goods are available. In [1,3], we prove that the corresponding mechanism is not only fair and Pareto optimal but also strategyproof. In addition, we address questions regarding computability and demonstrate how desired outcomes arise as limits of some natural proportional or best response dynamics for both models.

With my work, I hope to contribute to our general understanding of fairness and illuminate ways for a fair and efficient provision of public goods.

Representative Papers:

[1] Coordinating Charitable Donations (Working paper)
with F. Brandt, E. Segal-Halevi, and W. Suksompong

[2] Optimal Budget Aggregation with Single-Peaked Preferences (EC 2024)
with F. Brandt, E. Segal-Halevi, and W. Suksompong

[3] Balanced Donor Coordination (EC 2023)
with F. Brandt, E. Segal-Halevi, and W. Suksompong

[4] Funding Public Projects: A Case for the Nash Product Rule (Journal of Mathematical Economics 2022, WINE 2021)
with F. Brandl, F. Brandt, D. Peters, C. Stricker, and W. Suksompong

DANIEL HALPERN (Homepage, CV)

Thesis: Social Choice in the Modern Era: Navigating AI, Uncertainty, and Unprecedented Scale (2025)

Advisor: Ariel D. Procaccia, Harvard University

Brief Biography: I am a Ph.D. student in computer science at Harvard University, where I am funded by an NSF graduate research fellowship and a Siebel Scholarship. Before joining Harvard, I completed a Bachelor's of Science at the University of Toronto, where I worked with Nisarg Shah.

Research Summary: Broadly, my research applies tools from social choice theory and fair division to new contexts, often inspired by artificial intelligence. In these settings, classical frameworks frequently fall short, because traditional assumptions no longer hold. To address these challenges, I develop new theoretical models to design provably robust systems, which I empirically validate on real data, when possible. Below, I highlight three specific instances of this overarching agenda:

In [1], we consider the problem of fine-tuning a Large Language Model (LLM), improving its outputs using human preference data between different prompt answers. This is inherently a social choice problem, as we must aggregate heterogenous human preferences into a single output LLM. However, to facilitate training, the output here must be a reward function, which can assign a score to an arbitrary LLM output. This does not fit neatly into any existing social choice framework which typically output a single answer or set of answers. Nevertheless, variants of axioms from social choice can still apply. We show that current aggregation methods fail to satisfy fundamental social choice properties, and complement this with new aggregation rules to address these issues.

In [2], we take the perspective of an opinion aggregation website such as pol.is, which facilitates a large-scale discussion on complex issues, from how to regulate climate change to how to prioritize city funds. Participants express preferences on statements submitted by others, and then the platform generates a summary of the opinion space. At first glance, this again resembles a social choice problem: aggregate participant preferences over statements into a summary. However, due to the platform's scale, it is infeasible to ask each participant for preferences on all possible statements. Thus we must make do with partial data. We theoretically demonstrate the representation properties we can guarantee and validate these findings using real data from Polis discussions.

In [3], we explore a novel form of governance called Liquid Democracy, made possible by modern digital platforms. Through both theoretical models and real human experiments in classroom settings, we demonstrate its strong performance at improving collective decision-making.

Representative Papers:

[1] Axioms for AI Alignment (NeurIPS'24 Spotlight Presentation) with L. Ge, E. Micha, A.D. Procaccia, I. Shapira, Y. Vorobeychik, J. Wu

[2] Representation with Incomplete Votes (AAAI'23) with G. Kehne, A.D. Procaccia, J. Tucker-Foltz, and M. Wüthrich

[3] Tracking Truth in Liquid Democracy (*Management Science*, EC'23) with A. Berinsky, J.Y. Halpern, A. Jadbabaie, E. Mossel, A.D. Procaccia, and M. Revel

MEENA JAGADEESAN (Homepage, CV, Google Scholar)

Thesis: Machine Learning Ecosystems of Self-Interested Agents ('25)

Advisor: Michael I. Jordan and Jacob Steinhardt, UC Berkeley

Brief Biography: I am a 5th year PhD student in Computer Science at UC Berkeley, affiliated with the Berkeley AI Research Lab. My research is supported by an Open Philanthropy AI Fellowship and a P.D. Soros Fellowship. I've interned at Microsoft Research in the Economics & Computation Group (summers '23, '24).

Research Summary: Modern ML models (e.g., LLMs and recommender systems) interact with humans, companies, and other models within a broader ecosystem. However, interactions between these agents often lead to unintended ecosystem-level outcomes, including clickbait, safety violations, and market concentration.

My research takes an economic perspective of these multi-agent interactions, towards a vision of ML ecosystems operating as well-functioning markets. I view models, humans, and companies as self-interested agents optimizing their own objectives. I aim to characterize how multi-agent interactions shape ecosystem-level outcomes, and develop interventions to steer outcomes towards societal objectives.

In LLM ecosystems, the companies which train or fine-tune LLMs compete for user usage. My research demonstrates how this form of competition distorts model performance and market structure. Specifically, when companies fine-tune the same LLM, we show that training the LLM with more resources can reduce user welfare [1]. Furthermore, when companies train different LLMs, we show that new companies can enter the market with much less data than incumbents [4]. The underlying driver is that companies strategically train their models to attract users.

On recommendation platforms, the ML model used for recommendations facilitates competition between users. My research demonstrates how user incentives amplify the impact of details of the ML model. Specifically, for content recommendation, we characterize how the recommendation model's learned embeddings shape the supply of available content, due to creator incentives [2]. Moreover, for matching platforms with prices, where stability captures user incentives, we design bandit-based recommendation algorithms minimizing cumulative instability [3].

In other work, I also leverage an economic perspective of ML ecosystems to study human-AI interactions, competing platforms, AI policy, and algorithmic fairness.

Representative Papers:

[1] Improved Bayes Risk Can Yield Reduced Social Welfare Under Competition (NeurIPS 2023) with M. I. Jordan, J. Steinhardt, and N. Haghtalab

[2] Supply-Side Equilibria in Recommender Systems (NeurIPS 2023) with N. Garg and J. Steinhardt

[3] Learning Equilibria in Matching Markets from Bandit Feedback (Full version at Journal of the ACM; conference version at NeurIPS 2021) with A. Wei, Y. Wang, M. I. Jordan, and J. Steinhardt

[4] Safety vs. Performance: How Multi-Objective Learning Reduces Barriers to Market Entry (Under submission) with M. I. Jordan and J. Steinhardt

DEVANSH JALOTA (Homepage, CV, Google Scholar)

Thesis: Algorithm and Incentive Design for Sustainable Resource Allocation ('25)

Advisor: Marco Pavone & Yinyu Ye, Stanford University

Brief Biography: Devansh is a final-year Ph.D. candidate in Computational and Mathematical Engineering at Stanford University, where he is a Thomas C. Nelson Stanford Interdisciplinary Graduate Fellow. Prior to joining Stanford, Devansh received his B.Sc. in Civil and Environmental Engineering and B.A. in Applied Mathematics from UC Berkeley.

Research Summary: Devansh's research develops data-driven and online learning algorithms and incentive schemes to advance the science and practice of market design for sustainable and society-aware resource allocation. Blending ideas from operations, economics, and computer science, his research pushes the frontiers of resource allocation through both *foundational* and *application-driven* research.

On the foundational front, he introduces and studies models that incorporate *societal* and *practical* considerations, including equity, fairness, privacy, uncertainty, and security, into classical resource allocation problems and develops foundational tools and algorithms for decision-making in these more complex settings. For instance, his work has addressed the privacy and information availability concerns of traditional equilibrium pricing approaches that rely on complete information of user preferences in the context of electricity and Fisher markets [1, 4]. Moreover, his work has laid the foundations for accommodating complex constraints arising due to fairness and security concerns in classical equilibrium models and developed algorithms for computing equilibria in these more complex settings [2, 5].

On the applied front, he leverages theory into applications, tailoring AI and optimization-driven algorithms for domains spanning future mobility systems [3], artificial currency markets [1, 2], and electricity markets [4]. Notably, his work on future mobility systems develops methods to address emerging transportation equity and data privacy and uncertainty challenges, with a keen focus on addressing the inequity issues surrounding road congestion pricing [3].

Representative Papers:

[1] Stochastic Online Fisher Markets: Static Pricing Limits and Adaptive Enhancements (Operations Research (Forthcoming), WINE 2023) with Y. Ye

[2] Fisher Markets with Additional Linear Constraints: Equilibrium Properties and Efficient Distributed Algorithms (Games and Economic Behavior 2023) with M. Pavone, Q. Qi, and Y. Ye

[3] Balancing Fairness and Efficiency in Traffic Routing via Interpolated Traffic Assignment (Journal of Autonomous Agents and Multi-agent Systems 2023) with K. Solovey, M. Tsao, S. Zoepf, and M. Pavone

[4] Online Learning for Equilibrium Pricing in Markets under Incomplete Information (Major Revision at Operations Research) with H. Sun, N. Azizan

[5] When Simple is Near-Optimal in Security Games (Working Paper) with M. Ostrovsky, M. Pavone

ANAND KALVIT (Homepage, CV)

Thesis: Improved Asymptotics for Multi-armed Bandit Experiments under Optimism-based Policies: Theory and Applications (2023)

Advisor: Assaf Zeevi, Columbia University Graduate School of Business

Brief Biography: Anand Kalvit is a postdoctoral fellow at Stanford University's Immigration Policy Lab, where he is exploring incentive-aware online learning approaches for algorithmic refugee assignment. He completed his PhD in Decision, Risk, and Operations from Columbia University's Graduate School of Business and holds Bachelor's and Master's degrees in Electrical Engineering from IIT Mumbai, India. In Fall 2023, he was a research fellow at the Simons Laufer Mathematical Sciences Institute (formerly MSRI), Berkeley. His work has been recognized with spotlight papers at NeurIPS 2021 and the INFORMS RMP Conference 2022, along with finalist placements in multiple INFORMS competitions.

Research Summary: Anand's research focuses on sequential decision-making under uncertainty, blending theory and applications at the intersection of online learning, optimization, and mechanism design. A central theme in his current work is guiding agents toward beneficial actions in dynamic systems, with applications ranging from recommender systems to refugee allocation.

His doctoral thesis delves into the analysis of multi-armed bandit algorithms using the diffusion approximation framework and the design of optimal policies for complex settings such as dynamic marketplaces, aiming to address practical challenges and translate theoretical insights into real-world solutions.

Looking ahead, Anand aims to integrate and advance methods from machine learning, optimization, and economics into societally impactful domains such as AI for social good. His ongoing projects explore adaptive queueing and personalized recommendations in healthcare contexts, with the goal of developing robust frameworks for decision-making in non-stationary and resource-constrained environments.

Representative Papers:

[1] Incentivized Exploration via Filtered Posterior Sampling (EC'24)
with Y. Gur, and A. Slivkins

[2] Complexity Analysis of a Countable-armed Bandit Problem (ALT'23)
with A. Zeevi

[3] Dynamic Learning in Large Matching Markets (NeurIPS'22)
with A. Zeevi

[4] Bandits with Dynamic Arm-acquisition Costs (Allerton'22)
with A. Zeevi

[5] A Closer Look at the Worst-case Behavior of Bandit Algorithms (NeurIPS'21)
with A. Zeevi

[6] From Finite to Countable-armed Bandits (NeurIPS'20)
with A. Zeevi

STANISŁAW KAŹMIEROWSKI (Homepage, CV)

Thesis: Solving Succinct Games ('25)

Advisor: Marcin Dziubiński, University of Warsaw

Brief Biography: I am a fourth-year PhD candidate at the University of Warsaw, Faculty of Mathematics, Informatics, and Mechanics, where I work on problems related to solving large games with succinct representation. During my PhD, I enjoyed a four-month-long internship at the Department of Economics of the University of Zurich, where I collaborated with Prof. Christian Ewerhart.

Research Summary: My research focuses on game theory, with a particular emphasis on the computation of equilibria in large games with succinct representations. I develop efficient algorithms to compute Nash equilibria in games with large, discrete strategy spaces, such as conflicts with multiple battlefields and network-based attack-defense games. A central challenge in these areas is the exponential growth in the number of strategies, where traditional methods often prove inefficient, and this is where my work seeks to innovate.

Beyond the computational aspect, I am also interested in the theoretical properties of equilibria. In our work on the Arad-Rubinstein game [3], we investigate how changing the tie-breaking rule affects the equilibrium set, revealing insights into strategic behavior, inefficiencies, and robustness.

To address the challenges posed by large games, I employ techniques such as strategy symmetrization, algorithmic optimization, and heuristic methods. For example, in article [1], we describe a network reduction operation that allows us to compute a Nash equilibrium in the Attack and Defense Game on Networks in polynomial time with respect to the number of nodes. In article [2], we present a polynomial-time algorithm for computing symmetrized payoffs in symmetric conflicts with multiple battlefields, reducing the game's size exponentially with a polynomial time cost. When combined with the Double Oracle Algorithm and a heuristic that leverages the model's structure, this method achieves a speedup of four orders of magnitude compared to classical approaches.

In my ongoing work (working single-author paper), I explore a variant of the Colonel Blotto game that incorporates costs, demonstrating that it is strategically equivalent to a zero-sum Colonel Blotto game with one additional battlefield. This equivalence allows for the efficient computation of Nash equilibria in polynomial time with respect to the total number of battlefields and resources available to the players.

Representative Papers:

[1] Computation of Nash Equilibria of Attack and Defense Games on Networks (SAGT 2023) with M. Dziubiński

[2] Efficient Method for Finding Optimal Strategies in Chopstick Auctions with Uniform Objects Values (AAMAS 2024) with M. Dziubiński

[3] An equilibrium analysis of the Arad-Rubinstein game (Journal of Economic Behavior & Organization) with C. Ewerhart

POOJA KULKARNI (Homepage, CV)

Thesis: Fair Allocation of Indivisibles Beyond Additive Valuations (2025)

Advisor: Ruta Mehta, Jugal Garg, University of Illinois at Urbana-Champaign

Brief Biography: I am a final-year Computer Science Ph.D. student at the University of Illinois at Urbana-Champaign (UIUC). My research focuses on Fair Allocation of Resources across various settings, including discrete, continuous, offline, and online environments. I have published in top theory conferences such as SODA, ITCS, ICALP, and AI conferences like AAAI and AAMAS, with much of my work involving submodular and XOS maximization. Before my Ph.D., I completed my Master's at Indian Institute of Science (IISc) and Bachelor's at College of Engineering Pune (CoEP), earning a gold medal at both institutions. During my PhD, I have interned at Meta and was offered an internship at Google (declined).

Research Summary: Resource allocation, such as in food banks or ad-slot allocation, requires fairness to maintain societal harmony. Fairness has been studied extensively, leading to various notions. My research addresses two key questions: (1) Do fair allocations exist? (2) Can they be computed efficiently? I study these in both offline settings, where agents, goods, and preferences are known in advance, and online settings, where agents or goods arrive over time.

Offline In the offline setting, my research focuses on agents with submodular and XOS preference functions. I've studied fairness notions like Nash Social Welfare (NSW) for submodular [1] and XOS valuations [2], as well as Any Price Share (APS) for submodular valuations [3] and other subclasses beyond additive. My work involves techniques such as (1) Linear and concave programming, and (2) Combinatorial methods like greedy and local search, exposing me to concepts like continuous extensions of submodular functions, configuration LPs, and market-based fairness approximations. In an upcoming paper, we introduce a new dependent rounding scheme for submodular allocations.

Online We study a setting where goods are known in advance, but agents arrive over time. Fairness is challenging in offline settings and becomes more complex with unpredictable future demands. While comparing to a prophet often yields strong negative results, we can address this in two ways: (1) Using alternative benchmarks and (2) Leveraging predictive information. Given the advances in learning-augmented algorithms, my upcoming work goes for the second solution. We give positive results for MMS-fair share allocations using modest predictions about agent arrivals. My goal is to characterize the trade-off between the amount (and cost) of information learned and the achievable fairness approximation.

Representative Papers:

[1] Approximating Nash social welfare under Submodular Valuations through (Un)matchings (SODA, TALG) with Jugal Garg, Rucha Kulkarni

[2] Sublinear Approximation for Nash social welfare with XOS Valuations (ITCS) with Siddharth Barman, Anand Krishna, Shivika Narang

[3] $\frac{1}{2}$ Approximate MMS Allocation for Separable Piecewise Linear Concave Valuations (AAAI) with Chandra Chekuri, Rucha Kulkarni, Ruta Mehta

TAO LIN (Homepage, CV)

Thesis: Incentives and Learning in Information Design ('25)

Advisor: Yiling Chen, Harvard University

Brief Biography: Tao Lin is a 5th-year PhD student in Computer Science at Harvard University. He obtained a BSc in EECS from Peking University. During his PhD, Tao interned at ByteDance and Google, and received a Siebel Scholarship.

Research Summary: As machine learning algorithms increasingly shape real-world decision-making, the *strategic behavior* of participating agents – whether users or data providers – fundamentally impacts the algorithmic performance. The design of learning algorithms in strategic, dynamic multi-agent environments departs from the traditional machine learning paradigm that assumes exogenous, stationary data distributions. I investigate the complex interplay between incentives and learning in both theoretical economic models and real-world multi-agent systems, contributing to the community's common goal of building socially responsible AI systems.

- *Incentives and learning in mechanism design* [1]: I study the fundamental question of equilibrium convergence in repeated auctions with learning agents. Internet ad auctions provide a canonical example, where advertisers employ online learning algorithms to bid for ad slots. While convergence in truthful auctions is known, the dynamics in non-truthful auctions remained an open challenge. My work [1] provides the first complete characterization of when mean-based learning algorithms converge to Nash equilibrium in repeated first-price auctions and when they do not.

- *Incentives and learning in information design* [2, 3]: Traditional information design (Bayesian persuasion) assumes that agents can optimally process received information through Bayesian updating – an assumption that rarely holds in practice. My research [2] develops a novel framework where agents instead learn from experience using no-regret algorithms. Surprisingly, our result shows that the optimal utility of the sender remains nearly unchanged when facing learning agents, compared with the traditional model. Moreover, this finding generalizes to any general principal-agent problems including Stackelberg games and contract design.

- *Empirical study: recommender systems* [4]: During an internship at ByteDance, I conducted research on the impact of content creators' incentives on the polarization phenomenon in recommender systems. This work demonstrates how theoretical insights about strategic behavior can inform the design of deployed AI systems, advancing the goal of building healthy and sustainable online information ecosystems.

Representative Papers:

[1] Nash Convergence of Mean-Based Learning Algorithms in Auctions (WWW'22) with X. Deng, X. Hu, W. Zheng

[2] Generalized Principal-Agent Problem with a Learning Agent (working paper) with Y. Chen

[3] Multi-Sender Persuasion: A Computational Perspective (ICML'24) with S. Hossain, T. Wang, Y. Chen, DC. Parkes, and H. Xu

[4] User-Creator Feature Dynamics in Recommender Systems with Dual Influence (NeurIPS'24) with K. Jin, A. Estornell, X. Zhang, Y. Chen, Y. Liu

ANDREAS MAGGIORI (Homepage, CV, Google Scholar)

Thesis: Beyond Worst-Case Analysis, With or Without Predictions ('23)

Advisor: Ola Svensson, Rudiger Urbanke, EPFL

Brief Biography: I am a Postdoc at the Data Science Institute of Columbia University, working with Eric Balkanski and Will Ma. I earned my PhD from EPFL, and during that time, I interned twice at Google Research, hosted by Nikos Parotsidis and Ehsan Kazemi, respectively.

Research Summary: My research focuses on the field of *decision-making under uncertainty*–essentially, how to optimize a function when only partial information about its input is available. Traditional models of uncertainty assume that inputs are either adversarial (worst-case analysis) or drawn from a known distribution (stochastic analysis). Both models have limitations: worst-case analysis is often too pessimistic, ignoring useful information like typical instances and historical data, while stochastic analysis lacks robustness against noise, corruptions, outliers, and distribution shifts.

Learning-Augmented Algorithms ([1], [2], [3]): Hence, a central focus of my research is on intermediate models, particularly the area of learning-augmented algorithms. In this framework, we assume access to potentially imperfect predictions and seek to leverage these predictions without making any assumptions about their quality. Specifically, if the predictions are accurate, the algorithm should perform near optimally; if the predictions are poor, the algorithm must maintain robustness. This approach stems from the reality that machine learning models, while often accurate in practice, rarely offer worst-case guarantees.

However, relying on predictions can be a double-edged sword, as the algorithms' performance can be highly sensitive to bias present in those predictions.

Fairness [3], [4]: A recent theme in my work examines the fairness implications of optimization under uncertainty, especially in online decision-making. In [3], we prove that learning-augmented algorithms can be particularly vulnerable to small biases in predictions and lead to very unfair outcomes. To overcome this limitation, we design an algorithm that can use biased data to ameliorate its performance while making fair decisions.

Representative Papers:

[1] Learning Augmented Energy Minimization via Speed Scaling (Spotlight at NeurIPS 2020) with E. Bamas, L. Rohwedder, and O. Svensson

[2] The Primal-Dual Method for Learning Augmented Algorithms (Oral at NeurIPS 2020) with E. Bamas, and O. Svensson

[3] Fair Secretaries with Unfair Predictions (to appear at NeurIPS 2024) with E. Balkanski, and W. Ma

[4] Fair and Consistent Correlation Clustering (Under Submission, 2024) with E. Balkanski, and I. Chatzitheodorou

DIVYARTHI MOHAN (Homepage, CV)

Thesis: Simplicity and Optimality in Algorithmic Economics: Multi-Item Auctions and Social Learning ('21)

Advisor: S. Matthew Weinberg, Princeton University

Brief Biography: Divyarthi Mohan is a postdoctoral researcher at Boston University hosted by Prof. Kira Goldner. Previously, she was a postdoc at Tel Aviv University with Prof. Michal Feldman. She obtained her PhD in Computer Science at Princeton University in July 2021 advised by Prof. Matt Weinberg. Divya was awarded the class of 2021 Siebel Scholarship and the Simons-Berkeley Research Fellowship for Fall 2022. During her PhD, she received the School of Engineering and Applied Science's Award for Excellence in 2019 and the Department of Computer Science's Graduate Student Teaching Award in 2018.

Research Summary: My research addresses complex strategic behaviors and informational uncertainties in various settings of algorithmic economics, both by tackling fundamental problems in mechanism design and by developing/studying models for emerging applications and phenomena.

In particular, my expertise is in *multi-dimensional mechanism design* [1,4], where optimal mechanisms are often extremely complex, computationally intractable or even impossible without strong assumptions. My research tackles this through the algorithmic lens of approximation and designs simple, computationally efficient algorithms that are robust to strategic behavior.

My recent work explores settings with interdependencies in agents' values. The celebrated interdependent values model, awarded with the 2020 Nobel Prize in Economics, has been pivotal in studying auctions with more realistic representation of agent valuations, as they crucially rely on private information of others. I design simple algorithms and truthful mechanisms that guarantee a constant approximation to the optimal welfare, under additional informational challenges: namely, private valuation functions [1] or online arrival of agents [2].

In addition, my work in *social learning and strategic communication* investigates how and why (mis)information is propagated due to strategic interactions [3]. I develop and study simple models of communication with the goal of understanding prevalent social phenomena such as herding, polarization and echo chambers.

Representative Papers:

[1] Constant Approximation for Private Interdependent Valuations (FOCS 2023, Highlights Beyond EC 2024)
with A. Eden, M. Feldman, K. Goldner, and S. Mauras

[2] Optimal Stopping with Interdependent Values (EC 2024)
with S. Mauras and R. Reiffenhäuser

[3] Communication with Anecdotes (ITCS 2024)
with N. Haghtalab, N. Immorlica, B. Lucier, and M. Mobius

[4] Approximation Schemes for a Unit-Demand Buyer with Independent Items via Symmetries (FOCS 2019)
with P. Kothari, A. Schvartzman, S. Singla, S.M. Weinberg.

MATHIEU MOLINA (Homepage, CV, Google Scholar)

Thesis: Fairness and Sequential Decision-Making ('25)

Advisor: Vianney Perchet, ENSAE, Patrick Loiseau and Nicolas Gast, Inria

Brief Biography: Mathieu Molina is in the final year of his PhD in applied mathematics at the Inria FAIRPLAY team and ENSAE Paris, supervised by Patrick Loiseau, Vianney Perchet, and Nicolas Gast. He graduated from Mines ParisTech in 2021 and holds a master's degree in Artificial Intelligence, Systems, and Data from PSL University. His research interests lie at the intersection of sequential decision making, and fairness, in problems such as prophet inequalities, multi-armed bandits, auctions, and bipartite matching.

Research Summary: My research explores the impact of various constraints on decision-making in classical algorithmic settings, with a special emphasis on fairness and efficiency.

In my most recent work [1], I study a variant of an i.i.d. prophet inequality where a decision-maker competes with a prophet that selects the average of the two best items, instead of the maximum. We demonstrate that this modification sharply improves the competitive ratio from 0.745 to 0.966, and for the top ℓ items, the competitive ratio approaches 1 exponentially fast as ℓ grows. I am currently extending this work to analyze how the introduction of fairness penalties impacts the competitive ratio in this setting.

A key challenge in online decision-making with constraints is that these constraints may be uncertain and need to be learned over time. In my work [2], I study multi-armed bandits with covering constraints, where each arm must secure a minimum expected reward. We develop algorithms that optimally balance constraint violation and regret by choosing between pessimistic and optimistic constraint estimators. In [3] I work on an online allocation problem with a fairness penalty, but where the decision-maker lacks direct information about protected groups. In this setting, we allow the decision-maker to purchase data of varying quality and cost, and design an algorithm that balances fairness, efficiency, and data acquisition costs.

Additionally, I examine how fairness constraints impact the performance of established mechanisms. In my work on bipartite matching markets [4], I investigate the "price of fairness" in terms of utility loss. We show that under certain fairness constraints related to equality of opportunity, the worst-case utility loss is linear in the number of protected groups but independent of the size of the matching graph.

Representative Papers:

[1] Prophet Inequalities: Competing with the Top ℓ Items is Easy (SODA'25)
with N. Gast, P. Loiseau, V. Perchet

[2] Multi-Armed Bandits with Guaranteed Revenue per Arm (AISTATS'24)
with D. Baudry, N. Merlis, H. Richard, V. Perchet

[3] Trading-off price for data quality to achieve fair online allocation (NeurIPS'23)
with N. Gast, P. Loiseau, V. Perchet

[4] The Price of Fairness in Bipartite Matching (Working paper)
with R. Castera, F. Garrido-Lucero, S. Mauras, P. Loiseau, V. Perchet

PAOLA MOSCARIELLO (Homepage, CV)

Thesis: Redistricting with Endogenous Policies (2024)

Advisor: Leeat Yariv, Princeton University

Brief Biography: I am a Microeconomic theorist and a PhD candidate in the Economics Department at Princeton University. I specialize in applying tools from information design and optimal transport to topics in political economy and behavioral economics.

Research Summary: I am interested in a broad range of topics, including gerrymandering, committee decision making, school choice mechanisms, and experimental approaches to decision theory.

In my job market paper, "Redistricting with Endogenous Policies," I examine the interaction between partisan gerrymandering and the policy positions of candidates at the district level. I develop a model where a gerrymanderer partitions voters into equipopulous districts to maximize the expected number of districts won by one party. The key innovation is allowing candidates' positions to depend on the distribution of voters within a district, making voting behavior endogenous to the redistricting process itself. I solve the gerrymanderer's problem using tools from the optimal transport literature, mapping the redistricting problem to a Monge-Kantorovich transport problem. My findings show that optimal districts create a wedge between moderate and extreme opponents, encouraging the emergence of extreme candidates. By diluting the power of moderate voters, optimal redistricting generates a distribution of district representatives that has at least two modes, contributing to increased policy polarization.

My paper "Information Avoidance in School Choice," published in Games and Economic Behavior (2024), investigates how students' concerns about self-image can lead to strategic misreporting of preferences in school choice mechanisms.

I have several ongoing projects, coauthored with colleagues. For instance, my ongoing work, "Reputation in a Committee with Multiple Principals: The Case of the FOMC," examines how career concerns impact individual behavior and collective outcomes in the Federal Open Market Committee. Another example is a work titled "Caution in the Face of Complexity," which explores the interaction between complexity and ambiguity aversion in decision-making.

Representative Papers:

[1] Redistricting with Endogenous Policies (working paper)

[2] Information Avoidance in School Choice (Games and Economic Behavior, 2024)

[3] Caution in the Face of Complexity (work in progress)
with G. de Clippel, P. Ortoleva, and K. Rozen

[4] Reputation in a Commiette with Multiple Principals: The Case of the FOMC (work in progress)
with M. Iaryczower, and G. Lopez Moctezuma

ANIKET MURHEKAR (Homepage, CV, Google Scholar)

Thesis: Fairness, Efficiency, and Incentives in Allocation and ML Problems ('25)

Advisor: Jugal Garg and Ruta Mehta, University of Illinois at Urbana-Champaign

Brief Biography: Aniket is a PhD candidate in Department of Computer Science at UIUC, where he is advised by Jugal Garg and Ruta Mehta. He has been a Research Intern at Google Research (2024) and Adobe Research (2022). He is the recepient of the Mavis Future Faculty Fellowship, the Siebel Scholarship, and the IIT Bombay Academic Prize. He holds a B.Tech. in CS from IIT Bombay.

Research Summary: My goal is to *develop algorithmic solutions with provable guarantees on fairness, efficiency, and incentives* for problems of societal and industrial importance. My work uses ideas from economics and social choice to quantify fairness and efficiency, and leverages techniques from algorithm design, economics, and game theory to design solutions. My research has two main directions:

(1) *Investigating fundamental questions regarding the existence and computation of fair and efficient allocations.* In discrete fair division, envy-freeness up to any item (EFX) is regarded as a central notion of fairness. The existence of EFX allocations is one of the most fundamental and enigmatic open problems in fair division. For allocating chores to agents with additive preferences, the existence of EFX allocations is open even for $n = 3$ agents, and the best result in terms of approximation was the existence of $O(n^2)$-EFX allocations. In recent work [1], we prove that 4-EFX allocations of chores always exist, thus showing the first constant-factor approximation of EFX. Another important open problem is the existence of allocations of chores that are both fair (EF1; a relaxation of EFX) and efficient (Pareto-optimal). We showed that such allocations exist and can be efficiently computed for certain structured instances, e.g., for $n = 3$ agents [2].

(2) *Addressing issues of fairness and incentives in machine learning systems*, such as federated learning (FL), by using ideas from game theory and social choice. FL allows agents with individual datasets to collaborate and train a joint model. However, differences in data distributions can lead to unfair and inefficient outcomes, and collusion among agents. Moreover, data-sharing costs may disincentivize agents from sharing their data, leading to free-riding. To address these issues, we used ideas from social choice theory to develop an FL protocol which returns a model that is fair, efficient, and robust to coalitions [3], and designed a mechanism inspired from public goods economics whose Nash equilibria incentivize data-sharing [4].

Representative Papers:

[1] Fair Division of Indivisible Chores via Earning Restricted Equilibria (*under submission*) with J. Garg and J. Qin

[2] Weighted EF1 and PO Allocations with Few Types of Agents or Chores (*IJCAI* '24) with J. Garg and J. Qin

[3] Fair Federated Learning via the Proportional Veto Core (*ICML* '24) with B.R. Chaudhury, Z. Yuan, B. Li, R. Mehta, and A.D. Procaccia

[4] Incentives in Federated Learning: Equilibria, Dynamics, and Mechanisms for Welfare Maximization (*NeurIPS* '23) with Z. Yuan, B.R. Chaudhury, B. Li, and R. Mehta

MARIOS PAPACHRISTOU (Homepage, CV, Google Scholar)

Thesis: From Contagion to Stability: Insights into Network Dynamics, Resilience and Stability (2025)

Advisor: Jon Kleinberg, Cornell University

Brief Biography: I am a fifth-year PhD Candidate at Cornell University advised by Jon Kleinberg. I work on the economics of neteworks exploring their roles within large-scale social and information systems, and understanding their wider societal implications. My research has been supported by an Onassis Scholarship, a LinkedIn Ph.D. Fellowship, a Cornell Fellowship, a grant from the A.G. Leventis Foundation, a grant from the Gerondelis Foundation. I have also spent time in industry, and particularly at Twitter, and Microsoft Research.

Research Summary: In my research, I leverage tools from probability, statistics, algorithms, economics, and machine learning to study how we can make complex network systems more resilient to cascading failures, study decentralized decision-making under privacy, develop statistical network models, and study whether human-like behavior emerges in complex systems simulated by agents powered by LLMs.

Specifically, I am interested in information diffusion and contagion, and how to remediate it. Examples include centralized decision-making algorithms for *remediating network contagion* [1, 2, 3], and decentralized learning and d*ecision-making algorithms to ensure resilience in the presence of privacy risks* [4, 5]. Moreover, I am interested in the *structural characteristics of networks that promote or stop contagion* [6], as well as how information diffuses in complex systems where the agents are black-box models simulated by LLMs and whether *LLMs can simulate human-like complex behavior* [7].

Representative Papers:

[1] Allocating Stimulus Checks in Times of Crisis (WWW 2022)
with J. Kleinberg

[2] Dynamic Interventions for Networked Contagions (WWW 2023)
with J. Kleinberg and S. Banerjee

[3] Optimal Resource Allocation for Remediating Networked Contagions (submitted R&R to Management Science, 2024) with J. Kleinberg and S. Banerjee

[4] Group Decision-Making among Privacy-Aware Agents (under review at Operations Research, 2024) with M. A. Rahimian

[5] Differentially Private Distributed Estimation and Learning (IISE Transactions, 2024) with M. A. Rahimian

[6] Core-periphery Models for Hypergraphs (KDD 2022)
with J. Kleinberg

[7] Network Formation and Dynamics among Multi-LLMs (working paper, 2024)
with Y. Yuan

MANEESHA PAPIREDDYGARI (Homepage, CV)

Thesis: Designing Markets for Information - A Generalized Approach. ('25)

Advisor: Bo Waggoner, University of Colorado, Boulder

Brief Biography: I am a fifth year PhD candidate at CU Boulder advised by Prof. Bo Waggoner and work closely with Prof. Rafael Frongillo. My research encompasses contract theory, prediction markets, designing Automated Market Makers (AMMs) and the economics of blockchain. During my PhD, I have been fortunate to be hosted as an intern by Prof. David Pennock and by Ethereum Foundation, and honoured for [1] to be selected for Highlights Beyond EC 2024. Prior to this, I completed my Masters in Economics at Delhi School of Economics and Bachelor's in Computer Science from IIIT-Hyderabad.

Research Summary: The explosion of interest in collecting data to train large-language models reinforces the need for eliciting more focused information from agents when appropriate, termed *information elicitation*. While its widespread adoption took a back seat due to regulations, my research looks into how seemingly unrelated tools can be used to accomplish elicitation. My research goal is to further develop fundamental insights into fields that surround information elicitation and develop robust theory on Automated Market Makers (AMMs).

Prediction markets, a well-studied field of EconCS, elicit predictions about a future event by enabling trading securities. Our work [1] lays a foundational bridge connecting them to Constant Function Market Makers (CFMMs), a prominent type of AMMs prevalent in the trillion-dollar trading landscape of Decentralized Finance (DeFi). This connection highlights a significant correlation between market-making axioms and desirable information-elicitation axioms. This connection also opens up a rich area for future research, as the literature in both fields can inter-operate and evolve together.

A key innovation in DeFi allows other agents, called Liquidity Providers (LP), to provide liquidity in the market to enable trades in exchange for trading fees. We leverage our equivalence result [1] to introduce a general LP framework to prediction markets in [2] and develop further insights into multidimensional fee.

In the classic principal-agent moral hazard problem, i.e. contract theory, the actions of agents are hidden from the principal. In [3] we show that contracts can be implemented via *Proper Scoring Rules* and this hidden action can be revealed without loss of generality. This is a helpful tool when the principal wishes to learn and better incentivize actions across time periods.

Representative Papers:

[1] An Axiomatic Characterization of CFMMs and Equivalence to Prediction Markets (ITCS 2024) with R. Frongillo and B. Waggoner

[2] A General Theory of Liquidity Provisioning for Automated Market Makers (Working Paper) with A. Bhaskara, and R. Frongillo

[3] Contracts with Information Acquisition, via Scoring Rules (EC 2022) with B. Waggoner

SIDDHARTH PRASAD (Homepage, CV, Google Scholar)

Thesis: Mechanism Design and Integer Programming in the Data Age ('25)

Advisors: Maria-Florina Balcan & Tuomas Sandholm, Carnegie Mellon University

Brief Biography: I am a final-year PhD student in Computer Science at Carnegie Mellon University. My research interests span artificial intelligence, mechanism and market design, machine learning, and operations research, and I have interned at Google Research where I worked on recommender systems with Craig Boutilier. My research has been recognized by a best poster honorable mention at the 2024 Mixed Integer Programming (MIP) workshop, an oral presentation at NeurIPS 2022, and a spotlight at NeurIPS 2021. I received a B.S. in Math and CS from Caltech.

Research Summary: The thesis of my research to date is that high performance—*e.g.,* revenue, social welfare, run-time, memory, *etc.*—in marketplaces can only be fully realized via a synergy of interdisciplinary approaches in mechanism design, integer programming, and machine learning. My goal is to improve computation and economic design for society's various markets. My research spans the full spectrum of new models/concepts, theory, and practical implementation/experiments.

Within mechanism design, I have designed algorithms, modeled new learning paradigms, and invented new mechanism classes for various settings including two-part tariffs, combinatorial auctions, shrinking markets, and general multidimensional mechanism design. A highlight here is the first framework for integrating side information into mechanisms to boost revenue while preserving efficiency and incentives in general multidimensional settings like combinatorial auctions [3].

Within integer programming, I have (1) developed a comprehensive generalization theory for data-driven cutting plane configuration [1, 2] and (2) made foundational (non-learning-based) contributions to the theory and practice of cutting planes [4]. Our generalization theory unveils new mathematical structure in the branch-and-cut algorithm and the canonical class of Gomory cuts [2] and is validated through experiments that show the impact of data-dependent parameter tuning. In a working paper [4] for which I was awarded *Best Poster Honorable Mention* at the 2024 MIP workshop, we propose a new technique for strengthening cover cuts—cutting planes that are critical to solvers like Gurobi—and fix an error in the definitive paper on this topic from 2000. We derive conditions when our new cuts define *facets* of the integer hull, which is the gold standard for cuts, and validate their practical use via experiments. Our cuts deliver strong numerical properties and are currently being tested within FICO Xpress and Cardinal Optimizer.

Representative Papers:

[1] Sample Complexity of Tree Search Configuration: Cutting Planes and Beyond (NeurIPS'21 Spotlight) with M.-F. Balcan, E. Vitercik & T. Sandholm.

[2] Structural Analysis of Branch-and-Cut and the Learnability of Gomory Mixed Integer Cuts (NeurIPS'22 Oral) with M.-F. Balcan, E. Vitercik & T. Sandholm.

[3] Bicriteria Multidimensional Mechanism Design with Side Information (NeurIPS'23) with M.-F. Balcan & T. Sandholm.

[4] New Sequence-Independent Lifting Techniques for Cutting Planes and When They Induce Facets (MIP'24) with E. Vitercik, M.-F. Balcan & T. Sandholm.

NIDHI RATHI (Homepage, CV, Google Scholar)

Thesis: Algorithmic and Hardness Results for Fundamental Fair-Division Problems (2021)

Advisor: Siddharth Barman and Mrinal K. Ghosh, Indian Institute of Science (IISc), Bangalore, India

Brief Biography: Nidhi Rathi is a Lise Meitner postdoctoral research fellow at Max Planck Institute for Informatics (MPI-INF), Saarbrücken, Germany, hosted by Danupon Nanongkai and Kurt Mehlhorn. Nidhi recieved her Ph.D. in Mathematics from IISc, Bangalore, India, advised by Siddharth Barman and Mrinal K. Ghosh. She has received a Commendation Certificate by the CS department, IISc, for her excellent Ph.D. thesis, and prestigious IBM Ph.D. Fellowship and Lise Meitner Postdoctoral Fellowship. Before joining MPI-INF, she was a postdoctoral research fellow at Aarhus University, Denmark, hosted by Ioannis Caragiannis.

Research Summary: My broad research interests lie in the design and analysis of algorithms with a focus on problems inspired by *computational social choice* and *algorithmic game theory*. The central focus of my research is fair division, which explores how to allocate a set of items among agents with varying preferences in a way that all parties consider as *fair*. While multiple hardness results exist for the problem of finding fair/efficient cake divisions (allocating a divisible resource), my work bypasses these computational barriers by [1] identifying the broad class of instances specified by a unifying property of monotone likelihood ratios for which polynomial-time algorithms exist for envy-freeness and various notions of economic efficiency, [2] developing an efficient algorithm with a multiplicative approximation factor of 1/2 (currently, the best known). It is often not possible to achieve fairness and efficiency together and distributions over (deterministic) allocations is a typical way of achieving the existence of such solutions. My work shows that that the above problem belongs to the complexity class of PPAD [3].

Whether EFX allocations exist is a major open problem in fair division of indivisible goods. In recent works [4,5], we propose a potent relaxation of EFX, namely, epistemic EFX, and show that it exists for any number of agents with monotone valuations and can be computed in polynomial time for additive valuations.

In general, I aim to explore different concepts of "fairness" in various theoretical problems in algorithmic design.

Representative Papers:

[1] Fair Cake Division Under Monotone Likelihood Ratios (EC'19 and MOR'22)
with Siddharth Barman

[2] Fair and Efficient Cake Division with Connected Pieces (WINE'19)
with Siddharth Barman, Eshwar Ram Arunachaleswaran and Rachitesh Kumar

[3] On the Complexity of Pareto-Optimal and Envy-Free Lotteries (AAMAS'24)
with Ioannis Caragiannis and Kristoffer Arnsfelt Hansen

[4] New Fairness Concepts for Allocating Indivisible Items (IJCAI'23)
with Ioannis Caragiannis, Jugal Garg, Eklavya Sharma, & Giovanna Varricchio

[5] Epistemic EFX Allocations Exist for Monotone Valuations (submitted to AAAI'24)
with Hannaneh Akrami

ROJIN REZVAN (Homepage, CV, Google Scholar)

Thesis: Alternate Revenue Benchmarks: Approximation and Computation of simple vs. optimal in Multi-dimensional Bayesian settings (anticipated: 2025)

Advisor: Shuchi Chawla, University of Texas at Austin

Brief Biography: Rojin Rezvan is a final year PhD student at the University of Texas at Austin, advised by Shuchi Chawla. She received her masters degree from the University of Wisconsin-Madison. She is broadly interested in algorithmic game theory and mechanism design. More specifically, she has done research in multi-dimensional mechanism design, fairness in auctions and fair allocation. She is generally interested in the intersection of mechanism design and other fields such as fairness and decentralized systems.

Research Summary: One of the main focuses of my PhD is on the paradigm of "Simple vs. Optimal" in mechanism design for multi-dimensional settings. Multi-item mechanisms can have undesirable properties such as unbounded revenue, *lottery* options in the menu and super-additive pricing function. To circumvent these issues, there are two paths to take: 1) Make some assumptions, such as independence over item value distributions and the buyers' value functions, 2) Examine the validity of the benchmark. The approach we took in [1] and [2] was the latter.

Our proposal is to compare any *simple* mechanism we design to a more realistic benchmark, called "Buy-many". In this setting, it is assumed that each buyer can interact with the menu multiple times. This ensures that super-additive pricing will not happen. The main difference now is while optimal revenue may be unbounded, the gap between revenue of optimal simple mechanisms such as item pricing and optimal buy-many mechanisms is logarithmic in the number of items. In [1], we propose a structure necessary over the item values, with which we will get fine-grained results in terms of approximation and computation of the buy-many revenue via item pricing. In [2], we extend these results and definitions to multi-buyer setting.

I am also interested in algorithmic fairness. In [4], we ask: is it possible that certain allocation algorithms in ad auctions introduce unfairness to the allocations in addition to the data? The answer is yes: an algorithm that always allocates to the highest bidder, such as FPA, could potentially turn minor differences in bids to large differences in allocation. To circumvent the issue, we propose two different algorithms that ensure fairness, while losing a fraction of the optimal social welfare, or consequently revenue. Currently, I am working to extend this work to cases where the advertisers have budgets.

Representative Papers:

[1] Pricing Ordered Items (STOC 22) with S. Chawla, Y. Teng, C. Tzamos

[2] Buy-many Mechanisms for Many Unit-demand Buyers (WINE 23) with S. Chawla, Y. Teng, C. Tzamos

[3] Prophet Secretary Against the Online Optimal (EC 23) with P. Duetting, E. Gergatsouli, Y. Teng, and A. Tsigonias-Dimitriadis

[4] Individually Fair Auctions for Mutli-Slot Sponsored Search (Best student paper at FORC 22) with C. Chawla, N. Sauerberg

XIZHI TAN (Homepage, CV, Google Scholar)

Thesis: Learning-Augmented Mechanism Design ('25)

Advisor: Vasilis Gkatzelis, Drexel University

Brief Biography: Xizhi Tan is a fifth-year PhD student in Computer Science at Drexel University, advised by Prof. Vasilis Gkatzelis. She interned at Google Research during the summers of 2023 and 2024. Her work has received the Exemplary Theory Track Paper Award at EC 2024 as well as the Jay Modi Memorial Award. She was a finalist for the 2023 Meta Research PhD Fellowship.

Research Summary: Worst-case analysis has been the predominant method for mathematically evaluating algorithms in computer science. On the positive side, a worst-case guarantee provides a useful signal regarding the robustness of the algorithm. However, it can often lead to uninformative bounds or impossibility results that may not reflect the real obstacles that arise in practice. This is evident in the rapid progress of machine learning, which has produced highly effective algorithms, many lacking non-trivial worst-case guarantees.

Motivated by this discrepancy, a surge of recent research, known as "algorithms with predictions," aims to develop robust algorithms guided by machine-learned predictions that combine the *robustness* of worst-case guarantees with stronger performance when the predictions are *consistent* with the truth. Our work [1] applies this framework to economic systems and initiates the study of learning-augmented mechanism design in the presence of strategic agents.

We showcase the ubiquitous power of predictions in various societal and economic settings. In [1], we propose a mechanism for the strategic facility location problem that outputs the optimal solution when the prediction is correct, without sacrificing any worst-case guarantees. In the metric distortion problem, our voting rules reach the Pareto frontier of consistency and robustness [2]. Beyond social choice problems, predictions prove valuable in a variety of economic contexts. For example, [3] demonstrates how predictions help achieve improved revenue in online auction settings. In general combinatorial auction settings, our learning-augmented clock auction achieves nearly optimal welfare with correct predictions, which significantly improving upon the pessimistic $\log(n)$ worst-case bound of clock auctions, while still achieving the $\log(n)$ bound even if the prediction is arbitrarily bad [4].

Representative Papers:

[1] Learning-Augmented Mechanism Design: Leveraging Predictions for Facility Location (MOR '23, EC '22)
with P. Agrawal, E. Balkanski, V. Gkatzelis, and T. Ou

[2] Learning-Augmented Metric Distortion via (p, q)-Veto Core (EC '24)
with B. Berger, M. Feldman, and V. Gkatzelis

[3] Online Mechanism Design with Predictions (EC '24 Exemplary Theory Track Paper Award) with E. Balkanski, V. Gkatzelis, and C. Zhu

[4] Clock Auctions Augmented with Unreliable Advice (SODA '25)
with V. Gkatzelis and D. Schoepflin

ANISH THILAGAR (Homepage, CV)

Thesis: Practical Guarantees in Forecasting Competitions: Accuracy, Efficiency, and Approximate Truthfulness ('25)

Advisor: Rafael Frongillo and Bo Waggoner, CU Boulder

Brief Biography: I am a 5th year PhD student in the CS Theory Group at CU Boulder. I was previously an undergraduate at Caltech where I double majored in Math and Computer Science and did research in machine learning, quatum computing, and DNA computing. In between, I spent 2 years as a Software Engineer at Google where I helped build and launch AutoML (now Vertex AI).

My main research interest is learning from strategic agents, but I am generally interested in theoretical machine learning, game theory, and mechanism design.

Research Summary: My work has primarily focused on designing and analyzing winner-take-all forecasting competitions with experts (Kaggle, Good Judgement Project, etc.). In these settings, forecasters submit predictions about future events to a mechanism that then chooses a single winner after the event outcomes are realized. It is well known that traditional mechanisms are not truthful and instead give players an incentive to extremize their predictions, so there is no guarantee that the experts report their beliefs or that the chosen winner is actually good.

We show that forecasters can instead have an incentive to behave the opposite way and misreport by hedging their beliefs. However, by analyzing the conditions that drive these contrasting behaviors, we are able to show that the traditional mechanism will be approximately truthful (strongly limiting how much any expert will mis-report) under practically achievable conditions [4].

While some truthful mechanisms for the general setting are known, we show that they all require a large number of events to guarantee the chosen winner is actually good. Instead, we present a class of mechanisms that are approximately truthful (strongly limiting how much any expert will mis-report) but require far fewer events to guarantee a good winner is chosen [1]. Then, we show that standard measures of correlation do not capture the notion that matters in this setting, and instead introduce a new metric that is able to do so. Furthermore, we show that our class of approximately truthful mechanisms are robust to those correlations, the first such guarantee in the literature [3].

Additionally, I have spent some time working on elicitation and loss function design. We found the first consistent polyhedral loss for top-k classification, and show that previously used losses solve other interesting problems [2].

Representative Papers:

[1] Efficient Competitions and Online Learning with Strategic Forecasters (EC'21)
with R. Frongillo, R. Gomez, and B. Waggoner

[2] Consistent Polyhedral Surrogates for Top-k Classification and Variants (ICML'22)
with J. Finocchiaro, R. Frongillo, and E. Goodwill

[3] Forecasting Competitions with Correlated Events (2023)
with R. Frongillo, M. Lladser, and B. Waggoner

[4] A Strategic Analysis of Traditional Forecasting Competitions (2024)
with M. Monroe, M. Hsu, and R. Frongillo

ARTEM TSIKIRIDIS (Homepage, CV)

Thesis: Design and Analysis of Auctions: Algorithms and Incentives ('23)

Advisor: Vangelis Markakis, Athens University of Economics and Business

Brief Biography: I am a postdoctoral researcher at Centrum Wiskunde & Informatica (CWI), hosted by Guido Schäfer. Previously, I completed my Ph.D. in Computer Science at Athens University of Economics and Business, where I was advised by Vangelis Markakis.

Research Summary: I am broadly interested in algorithmic game theory, mechanism design, and online algorithms. During my PhD, I focused on the design and analysis of auctions, specifically on protocols that retain provable performance guarantees while being implementable in real-life scenarios. An example is [1], where we studied core-selecting mechanisms, a formalism introduced by Ausubel and Milgrom (2002). Although these auctions are not generally truthful, they offer strong revenue guarantees and align bidder incentives (in a weaker sense). Our contributions include identifying core-selecting mechanisms suitable for practical implementation and designing a prior-free truthful mechanism competitive with the minimum revenue in the core. I've also worked on budget-feasible mechanisms and equilibria in non-truthful auctions.

Recently, I have become interested in mechanism design in environments with predictions, also known as learning-augmented mechanism design. This beyond-worst-case analysis paradigm suggests augmenting algorithm inputs with predictions, leveraging potentially erroneous predictions to improve worst-case performance. The goal is to achieve strong performance when the prediction is perfect (consistency) while also providing guarantees when it's not (robustness). A recent line of work has proposed exploring strategic settings under this framework, a direction I find fascinating. In [2], for example, we study variants of the generalized assignment problem (GAP) in a setting without money. The main result is establishing the best possible consistency-robustness tradeoff for bipartite matching by designing a truthful mechanism that implements Gale-Shapley's deferred acceptance algorithm. Additionally, we design randomized mechanisms for more general GAP variants, achieving improved approximations compared to settings without predictions while maintaining a degree of robustness.

Another area I have worked on is stochastic optimization. In [3], we propose an extension of the Pandora's Box problem (Weitzman, 1979) that incorporates the notion of time in a general sense. For this NP-Hard problem, we provide an efficient constant-factor approximation to the optimal strategy of the decision maker.

Representative Papers:

[1] On Core-Selecting and Core-Competitive Mechanisms for Binary Single Parameter Auctions (WINE 2019) with E. Markakis

[2] To Trust or Not to Trust: Assignment Mechanisms with Predictions in the Private Graph Model (EC 2024) with R. Colini-Baldeschi, S. Klumper, and G. Schäfer

[3] Pandora's Box Problem Over Time (WINE 2024) with G. Amanatidis, F. Fusco and R. Reiffenhäuser

JAMIE TUCKER-FOLTZ (Homepage, CV)

Thesis: Algorithmic Institutional Fairness ('25)

Advisor: Ariel D. Procaccia, Harvard University

Brief Biography: I am a fifth year computer science PhD student at Harvard University. I earned my undergraduate degree in CS and mathematics from Amherst College, and a master's degree in CS from the University of Cambridge, where I studied on a Churchill Scholarship. At Harvard, I have been supported by both an NSF Graduate Research Fellowship and a Google PhD Fellowship.

Research Summary: I am broadly interested in applying ideas from theoretical computer science to improve political and economic institutions. I am particularly focused on algorithms for guaranteeing fairness in complex resource allocation tasks and democratic political processes. I work on a range of problems, either inspired by or directly applicable to real-world issues, drawing primarily upon methodologies from fair division, algorithmic game theory, and algorithmic techniques including combinatorial optimization, Markov chains, and computational geometry.

My first main research thread focuses on adapting benchmarks from the literature on fair division such as *proportionality* and *envy-freeness* beyond the realm of private goods to novel, high-stakes domains. For example, two of my recent papers have studied algorithms for political redistricting [2] and assigning school attendance zones [4], modeled as constrained fair division problems where the "players" are respectively political parties and demographic groups we wish to be fair toward. I have applied a similar axiomatic approach to the problem of randomized apportionment [3], refining the vast space of *ex ante proportional* algorithms by requiring that they are fair and predictable to arbitrary coalitions of parties.

My second research thread concerns graph algorithms for judging fairness in political redistricting, not via intrinsic fairness axioms but, rather, random sampling. If a given map gives some party far fewer congressional seats than a random map would, that is strong evidence it has been gerrymandered—this approach has been successfully used in high-profile litigation in the United States. There are numerous open technical questions regarding what we should mean by "random maps" and how we can efficiently sample them. My contributions include showing that the widely-studied *spanning tree distribution* favors geographically compact maps [5], and establishing the first polynomial-time algorithm to sample from it [1].

Representative Papers:

[1] Sampling Balanced Forests of Grids in Polynomial Time (STOC 2024)
with S. Cannon and W. Pegden

[2] You Can Have Your Cake and Redistrict It Too (EC 2023)
with G. Benadè and A.D. Procaccia

[3] Monotone Randomized Apportionment (EC 2024)
with J. Correa, P. Gölz, U. Schmidt-Kraepelin, and V. Verdugo

[4] School Redistricting: Wiping Unfairness Off the Map (SODA 2024)
with A.D. Procaccia and I. Robinson

[5] Compact Redistricting Plans Have Many Spanning Trees (SODA 2022)
with A.D. Procaccia

MARTIN VAETH (Homepage, CV)

Thesis: Essays in Information and Mechanism Design ('25)

Advisor: Roland Bénabou, Alessandro Lizzeri, and Fedor Sandomirskiy (Princeton University)

Brief Biography: I am currently in the sixth and final year of my PhD in Economics at Princeton University. Previously, I completed a MSc in Economics and Philosophy at the LSE and a BSc in Mathematics at Heidelberg University.

Research Summary: My research uses methods from information and mechanism design to tackle problems in bounded rationality and political economy.

My job market paper *Rational Voter Learning, Issue Alignment, and Polarization* [1] embeds costly information acquisition into a model of electoral competition to explain observed features of voter ideology. I model electoral competition between two parties when voters can learn about their political positions through flexibly acquiring costly information. Optimal learning creates *polarized* and *aligned* political preferences even when the true distribution of ideal points is unimodal and independent across policy issues. That is, voters' revealed positions are bimodally distributed and correlated across diverse issues such as taxation and abortion. When party positions are strategically chosen, voter and party polarization are mutually reinforcing, and both increase as information becomes less costly. This can explain the rise in polarization in recent decades in the US through the advance of information technologies like the internet.

My paper *Attention and Regret* [2] explores a connection between costly information acquisition/attention and emotions. The paper provides an evolutionary explanation for regret as an optimal self-control mechanism to motivate attention and thereby improve decision-making. The model endogenizes the optimal emotions as incentives for an agent who overweights the cost of attention, for example due to temptation or present bias. The optimal emotions turn out to follow the functional form of classical regret theory, which was proposed by Bell (1982) and Loomes and Sugden (1982) to explain behavioral anomalies. Further, the model advances regret theory by explaining why regret is stronger than rejoicing and why regret is stronger in simpler decision problems. Methodologically, the paper combines techniques from costly information acquisition with mechanism design.

In the paper *Imprecision Attenuates Updating* [3], I present a novel comparative statics result for Bayesian updating. Economists frequently model incomplete information through noisy normal signals about a normally distributed state. This signal structure can be used to explain behavioral inertia, as the posterior mean is compressed towards the prior mean, and this attenuation effect is stronger the less precise the signal. Despite the ubiquity of the normal-normal model, it was not known to what extent these properties generalize beyond normal distributions. I show that they generalize to all symmetric log-concave distributions.

Representative Papers:

[1] Rational Voter Learning, Issue Alignment, and Polarization (Under Review)

[2] Attention and Regret (R&R JPE)

[3] Imprecision Attenuates Updating (Working Paper)

GRIGORIS VELEGKAS (Homepage, CV, Google Scholar)

Thesis: Towards Addressing Challenges in Modern ML: Generalization and Responsible AI (2025)

Advisor: Amin Karbasi, Yale University

Brief Biography: I am a final-year PhD student in Computer Science at Yale University, advised by Amin Karbasi. I obtained my BSc and MSc in Electrical Engineering and Computer Science from NTUA. From May 2023 until October 2024 I was an intern at Google Research.

Research Summary: My research focuses on three main directions: **i)** understanding generalization properties of ML algorithms, **ii)** exploring responsible use of ML systems, and designing algorithms with provable replicability guarantees, **iii)** understanding the interaction between ML algorithms and mechanisms.

Modern ML systems are trained using neural networks with parameters vastly exceeding the amount of training data. Conventional theoretical analysis, based on the VC theory, suggests that such systems should suffer from *overfitting*. Thus, this worst-case analysis fails to explain their practical success. In [1] and [2] we derive *beyond* worst-case generalization guarantees, called *universal* rates, for two different learning tasks, interactive learning and regression. Under this notion of learnability, we derive much less pessimistic bounds than the VC theory suggests, that are more closely aligned with empirical observations in deep learning.

Another critical challenge is the replicability issue. The past decade has seen a *replicability crisis* across natural sciences, evident in ML. Many researchers struggle to replicate results from other studies and even their own. Thus, developing algorithms with *provable* replicability guarantees is crucial, ensuring that if executed twice on independent samples, results remain consistent. In [3], we develop replicable algorithms for learning large-margin halfspaces, a fundamental problem in learning theory, with *exponentially* smaller sample complexity than prior work.

Lastly, understanding how learning algorithms interact with economic mechanisms is equally important. Traditional analyses assume that entities interacting with mechanisms are *rational agents*. However, in applications like keyword auctions, *learning algorithms* bid on behalf of advertisers. Thus, it is crucial to study whether these analyses hold in their presence. In [4], we show that Myerson's celebrated auction is not optimal when bids are submitted by learning algorithms, proving that the revenue-optimal auction in this setting needs to be *randomized*.

Representative Papers:

[1] Universal Rates for Interactive Learning (NeurIPS 2022, Oral)
with S. Hanneke, A. Karbasi, and S. Moran

[2] Universal Rates for Regression: Separations between Cut-Off and Absolute Loss (COLT 2024) with I. Attias, S. Hanneke, A. Kalavasis, and A. Karbasi

[3] Replicable Learning of Large-Margin Halfspaces (ICML 2024, Spotlight)
with A. Kalavasis, A. Karbasi, K. G. Larsen, and F. Zhou

[4] Randomized Truthful Auctions with Learning Agents (NeurIPS 2024)
with G. Aggarwal, A. Gupta, and A. Perlroth

JEREMY VOLLEN (Homepage, CV, Google Scholar)

Thesis: Algorithms for Representation-based Fairness in Collective Decisions ('25)

Advisors: Haris Aziz and Toby Walsh, UNSW Sydney

Brief Biography: I am a Ph.D. candidate in my final year at UNSW Sydney. In Summer 2023, I visited Stanford University and was hosted by Ashish Goel. Previously, I received my Bachelor's in Computational Mathematics from Rice University.

Research Summary: My research focuses on the design of collective decision-making systems, with a particular focus on systems which guarantee fair outcomes. The central contributions of my work are two-fold: (1) novel and meaningful fairness definitions, which often take inspiration from social choice theory and fair division, and (2) efficient algorithms which advance the frontier of fair decision-making by provably guaranteeing the defined properties.

In even the simplest voting settings, deterministic algorithms encounter strong impossibilities when pursuing fair decisions. As seen by the universality of the coin toss, randomization is a natural approach to fairness in the face of these obstacles. This approach necessitates meaningful *ex-ante* fairness definitions. In the setting in which a collective must select k alternatives, we propose [1] a new definition which is stronger than all fairness properties known to admit efficient algorithms. We then use flow networks to design efficient algorithms which satisfy our definition in conjunction with other desiderata. One such algorithm computes lotteries over outcomes which additionally satisfy strong fairness guarantees *ex-post*. In [2], we extend this approach, known as "best-of-both-worlds fairness", to participatory budgeting (PB), a direct democratic process that allows participants to decide how to spend a central budget. In addition to designing fair algorithms, we develop a technique to implement arbitrary divisible outcomes by lotteries over discrete PB outcomes while minimizing the variance in budget spent ex-post.

My work also investigates contexts in which the use of techniques from computational social choice is relatively unexplored. One such work [3] studies the ubiquitous problem of centroid clustering. Motivated by scenarios in which data points correspond to individuals, we introduce properties which capture proportional representation in clustering outcomes. Our algorithms uphold the state-of-the-art with respect to existing fairness definitions while also providing proportional representation. In [4], we introduce a framework for PB-like processes for collectives that lack the institutional structure required to pool resources. We explore the extent to which systems can approximate welfare optimality while incentivizing participation.

Representative Papers:

[1] Maximum Flow is Fair: A Network Flow Approach to Committee Voting (EC 2024) with M. Suzuki

[2] Fair Lotteries for Participatory Budgeting (AAAI 2024) with H. Aziz, X. Lu, M. Suzuki, and T. Walsh

[3] Proportionally Representative Clustering (WINE 2024) with H. Aziz, B.E. Lee, and S. Morota Chu

[4] Coordinating Monetary Contributions in Participatory Budgeting (SAGT 2023) with H. Aziz, S. Gujar, M. Padala, and M. Suzuki

YONGZHAO WANG (Homepage, CV, Google Scholar)

Thesis: Multiagent Learning by Iterative Refinement of Game Models ('23)

Advisor: Michael Wellman, University of Michigan

Brief Biography: Yongzhao Wang is a postdoctoral fellow at the Alan Turing Institute in the United Kingdom hosted by Rahul Savani and Theodore Turocy. He is also an honorary research associate at the university of Liverpool, lecturing a second-year undergraduate course "Computer-Based Trading in Financial Markets" in Spring 2024 and Spring 2025. He received her Ph.D. in August 2023 from the CSE Department at the University of Michigan.

Research Summary: My broad and long-term research interests lie in Artificial Intelligence (AI) with a focus on game theory, reinforcement learning, and large language models in the study of multiagent systems. My work has focused on developing scalable and robust automated game-theoretic analysis frameworks for large complex multiagent systems, characterized by a large number of strategies and players, incomplete information across players, or sequential decision-making processes. Specifically, I have explored the following areas:

(1) *Learning in large and complex multiagent systems*: As the multiagent systems become large and complex, traditional game-theoretic analysis often struggles to handle. I integrated various learning methods (e.g., deep RL) in AI with game theory, leveraging AI's capabilities to improve the effectiveness of game-theoretic analysis in tackling these large, complex scenarios.

(2) *Large language models (LLMs) for multiagent systems*: Driven by the success of LLMs, the potential applications of LLMs in combination with game theory has become a promising new area of research. I investigated this combination from two directions: (1) integrating game-theoretic methods and software as modules within LLMs for automated game-theoretic analysis from natural language and (2) embedding LLMs within existing game-theoretic frameworks, such as simulating human behavior for the study of behavioral game theory.

(3) *Applications in cybersecurity and financial markets*: Multiagent systems are universal in the real world from entertainment games like poker to financial markets with millions of traders and cybersecurity scenarios involving attackers and defenders. In the financial domain, I applied game-theoretic analysis to understand various financial operations, market making, and market manipulation. In cybersecurity, I collaborated with domain experts to construct realistic game models and developed automated defense systems using AI and game theory.

Representative Papers:

[1] Market Making with Learned Beta Policies (ICAIF-24) with Rahul Savani, Anri Gu, Chris Mascioli, Theodore Turocy, and Michael Wellman.

[2] A Strategic Analysis of Prepayments in Financial Credit Networks (IJCAI-24) with Hao Zhou, Konstantinos Varsos, Nicholas Bishop, Rahul Savani, Anisoara Calinescu, and Michael Wooldridge.

[3] Evaluating Strategy Exploration in Empirical Game-Theoretic Analysis (AAMAS-22) with Michael Wellman.

MITCHELL WATT (Homepage, CV, Google Scholar)

Thesis: Designing Price Mechanisms for Large Markets

Advisor: Paul Milgrom, Stanford University

Brief Biography: I am a Ph.D. candidate in economics at Stanford University. I hold a Masters in Public Policy from Harvard Kennedy School and a Bachelor of Science in mathematics from the University of Queensland. Outside of academia, I have worked as a consultant at Auctionomics on online display advertising auctions and was a policy adviser and speechwriter for an Australian parliamentarian.

Research Summary: I study microeconomic theory and market design, with a focus on questions relevant to public policy and regulation.

My job market paper [1] and companion paper [2] study the design of in-kind subsidies for redistribution. In [1], we characterize the optimal subsidy mechanism when consumers can "top up" subsidized allocations in a private market. The ability to top up requires the planner to offer subsidies increasing in the consumer's demand for the good, limiting the scope of redistribution via subsidies. When the social planner seeks to redistribute to consumers with lower demand, subsidies are optimal only if lump-sum transfers are unavailable and the cost of public funds is lower than the average weight the planner assigns to consumer surplus, leading to subsidies for consumption up to a maximum level. When the social planner seeks to redistribute to consumers with higher demand, the social planner may prefer in-kind subsidies to lump-sum transfers, providing discounts for consumption beyond a minimum level. The optimal mechanisms have features of food stamps and fare capping programs observed in practice. In [2], we study the case in which consumers participate in either a private market or a subsidized program, but not both. This widens the scope of redistribution for the planner compared to the case with topping up, and the optimal mechanism has three components: a public option, nonlinear subsidies, and laissez-faire consumption.

In [3] and [4], I study Walrasian economies, relaxing in each case one of the standard assumptions of the classical model: convex preferences (in [3]) and price-taking (in [4]). In [3], we introduce Markup equilbrium, an extension of Walrasian equilibrium that adds a markup to the prices that consumers pay to ensure existence even in nonconvex quasilinear economies. Markup equilibria are resource-feasible, incur no budget deficit, and require little more communication and computation than the Walrasian equilibrium. In [4], I study the rate of convergence of price-taking incentives in the Walrasian model, showing that the price impact of misreports is inversely proportional to the number of agents (with high probability) when the expected demand correspondence is strongly monotone.

Representative Papers:

[1] In-Kind Subsidies with Topping Up (Job Market Paper) with Z. Y. Kang

[2] Optimal In-Kind Redistribution with Z. Y. Kang

[3] A Walrasian Mechanism with Markups for Nonconvex Economies (EC 2022, Revise and Resubmit at *Review of Economic Studies*) with P. Milgrom

[4] Strong Monotonicity and Perturbation-Proofness of Walrasian Equilibrium (Best Paper by Young Researcher at Econometric Society Australasian Meeting 2023)

JIBANG WU (Homepage, CV, Google Scholar)

Thesis: Strategic Alignment for AI Systems ('25)

Advisor: Haifeng Xu, University of Chicago

Brief Biography: Jibang is currently a final year PhD student in Computer Science at University of Chicago advised by Prof. Haifeng Xu. His research interest lies at the interface between game theory, learning theory and optimization, with the primary focus on modeling and solving multi-agent decision-making problems under complex, unknown environment. His work has recently received the Stigler center PhD dissertation award.

Research Summary: While intelligent systems are becoming more advanced and influential, their design often overlooks critical incentive structures within their operating environments, risking unintended and potentially harmful consequences. My research aims to advance the design principles and approaches of intelligent systems towards *strategic alignment*, a concept centered on aligning the interests of all stakeholders to achieve mutually beneficial outcomes. Examining the theoretical foundations of machine learning and algorithmic economics, my research branches into two key components of intelligent systems:

- Decision Alignment: *Learning for Strategic Decision-Making.* The outcomes of data-driven decisions can be subject to the strategic responses from stakeholders under conflicting interest and asymmetric information. My work [1,2] models the strategic interactions in the multi-agent decision-making processes and adopts the online learning framework to analyze the adaptive decision optimization problems in a complex, unknown environment.

- Feedback Alignment: *Learning from Strategic Data Sources.* The data empowering machine learning systems can be strategically withheld or manipulated by the data providers. My work [3,4] models data providers' incentives on the learning outcomes and adopts a mechanism design perspective to analyze how different design of statistical methods (e.g., ranking, classification or calibration) or monetary incentives could induce more desirable equilibrium outcomes.

Building upon the two threads above, my ongoing research agenda is focusing on strategic alignment problems in the emerging realm of generative AI. In particular, I am interested in developing practical techniques to 1) build *incentive-aware AI agents with strategic intelligence and rationalizable behaviors*; 2) align the economic incentives of users, model developers and data providers *for more sustainable AI ecosystems*.

Representative Papers:

[1] Markov Persuasion Process and Its Efficient Reinforcement Learning (EC 22)
with Z. Zhang, Z. Feng, Z. Wang, Z. Yang, M. I. Jordan, H. Xu

[2] Robust Stackelberg Equilibria (EC 23)
with J. Gan, M. Han, H. Xu

[3] Auctioning with Strategically Reticent Bidders (WINE 24)
with A. Badanidiyuru, H. Xu

[4] An Isotonic Mechanism for Overlapping Ownership (ongoing)
with H. Xu, Y. Guo, W. Su

BRIAN HU ZHANG (Homepage, CV, Google Scholar)

Thesis: New Solution Concepts, Algorithms, and Applications for Extensive-Form Games: Learning, Correlation, Communication, and Common Knowledge ('25)

Advisor: Tuomas Sandholm, Carnegie Mellon University (CMU)

Brief Biography: I am a final-year PhD student in Computer Science at CMU. I have been honored to receive the *inaugural* CMU Hans J. Berliner Graduate Fellowship in AI (awarded to one student annually) for my research.

Research Summary: My research focuses on algorithms and solution concepts for solving large imperfect-information games. I have developed *new state-of-the-art and provably optimal* algorithms for many zero-sum and general-sum games. Examples include the *first solution concept and efficient algorithm* for hidden-role or social deduction games [1], the *first provably-optimal parameterized algorithm* for optimal (*e.g.,* welfare-maximizing) extensive-form correlated equilibria [2], and the *first efficient no-regret learning algorithms* for achieving the strongest robustness notions known to be efficiently achievable [3].

A foundation that enabled these results (and more!) stems from intrinsic *connections* that I have discovered among problems previously treated as unrelated. For example, the new no-regret algorithms above were made possible in part by exploiting new insights from the seemingly-unrelated problem of generalized mechanism design [3]. These connections are part of a *new framework* I have developed that 1) unifies under a single umbrella a growing range of problems such as principal-agent problems (including mechanism, information, and contract design) and optimal extensive-form correlated equilibria, and 2) *reduces them all to zero-sum games* (two-player and team games, respectively). This allows them to be solved with the rich zero-sum game toolbox—including ML techniques—and motivates further advances to the state of the art in zero-sum games [5].

My research has also led to *practical breakthroughs*. I have created the *first superhuman AI for dark chess* [4], the most complex turn-based game in which superhuman AI has been achieved. Further, my new algorithm and solution concept for hidden-role games led to the *first exact solutions to several variants of the popular game Avalon* [1], the most complex hidden-role game ever solved.

Representative Papers:

[1] Hidden-Role Games: Equilibrium Concepts and Computation (EC'24) with L. Carminati, G. Farina, N. Gatti, & T. Sandholm

[2] Optimal Correlated Equilibria in General-Sum Extensive-Form Games: Fixed-Parameter Algorithms, Hardness, and Two-Sided Column-Generation (EC'22; accepted to appear in Math of OR) with G. Farina, A. Celli, & T. Sandholm

[3] Efficient Φ-Regret Minimization with Low-Degree Swap Deviations in Extensive-Form Games (NeurIPS'24) with I. Anagnostides, G. Farina, & T. Sandholm

[4] Superhuman Performance in Dark Chess via General-Purpose Search Techniques in Imperfect-Information Games (Working paper'24) with T. Sandholm

[5] Computing Optimal Equilibria and Mechanisms via Learning in Zero-Sum Extensive-Form Games (NeurIPS'23) with G. Farina, I. Anagnostides, F. Cacciamani, S. McAleer, A. Haupt, A. Celli, N. Gatti, V. Conitzer, & T. Sandholm

JIAYU (KAMESSI) ZHAO (Homepage, CV)

Thesis: Incentivizing Flexibility in Platform Operations ('25)

Advisor: Daniel Freund, MIT

Brief Biography: I am a final year PhD student at the Operations Research Center at MIT, where I am advised by Prof. Daniel Freund. Prior to my PhD, I graduated Summa Cum Laude from Columbia University in 2020 with a B.S. degree in Operations Research.

Research Summary: My research studies how two-sided service platforms, via market and algorithm designs, can incentivize agents' flexibility to enhance operational efficiency. Such incentives facilitate the matching between supply and demand sides of the market by easing the heterogeneity in space (e.g., Lyft's relocation incentives for drivers), time (Uber's 'wait and save' discount for riders), among others. Motivated by the increasing uses of such flexibility incentives today, my research studies the design of flexibility in platforms, using a combination of tools from stochastic decision-making and game theory.

The first part of my research agenda studies the market design questions around flexible operations. While flexibility incentives are common on both the demand (e.g., 'wait and save' feature at Uber) and the supply side (Ride streak bonuses at Uber) of platforms, they have been treated *in isolation* in the literature and in practice. By modeling how these incentives influence the likelihood of compatibility between agents and the resulting matching size, my work [1] is the first to investigate the management of two-sided flexibility in platforms. Aside from this horizontal interplay of flexibility incentives on different market sides, my research also investigates the vertical supply chain implications of ride-hailing platforms' flexibility decisions [2]. When dual-sourcing autonomous vehicles (AVs) and flexible human drivers with self-scheduling capacity, platforms (e.g., Uber's operations in Phoenix) make dispatch prioritization decisions to fulfill demand through a hybrid fleet, which affects the incentives of AV suppliers and human drivers. I study how potential incentive misalignment can hinder successful AV deployments and provide contracting solutions to overcome them.

The second aspect of my research focuses on algorithms that provide *better customization and timing* to harness flexibility. For instance, booking platforms can adjust their admission control decisions in real-time by considering customers' heterogeneous probabilities of being no-shows (i.e., not requiring service) and their compensation requirements for overbooking. I study an online resource allocation problem that allows overbooking in [3] and propose a policy that improves the additive profit loss guarantee (compared to a clairvoyant) in T periods from $\Omega(\sqrt{T})$ in the literature to $\mathcal{O}(1)$ in our paper.

Representative Papers:

[1] Two-sided flexibility in platforms. (MIT ORC Best Student Paper Award) with D. Freund, and S. Martin

[2] On the supply of autonomous vehicles in platforms (EC'24) with D. Freund, and I. Lobel

[3] Overbooking with bounded loss. (EC'21, Mathematics of Operations Research) with D. Freund

Index

Designing Choice Architecture to Mitigate Selection Bias in Consumer Data Sharing

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

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

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

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

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

1. INTRODUCTION

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4. MAIN FINDINGS

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

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

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

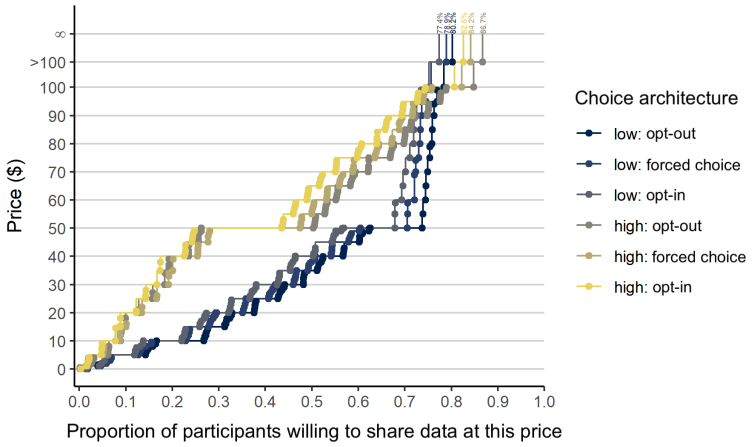


Fig. 1: Data supply curves across choice architecture treatments

decreases data valuation by 52.6% on average.

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

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

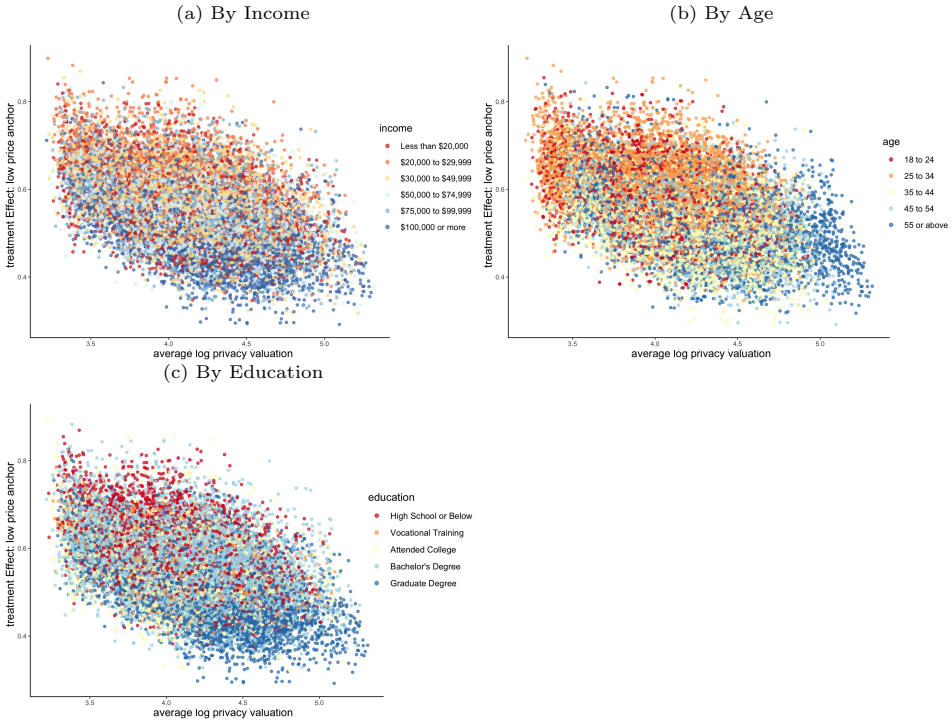
To illustrate this point empirically, we focus on demographic subgroups where such a negative correlation exists between privacy valuation and choice architecture effects. In our setting, younger, poorer, and less educated consumers tend to value their personal data less, but are also more responsive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

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

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

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

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



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

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

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

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

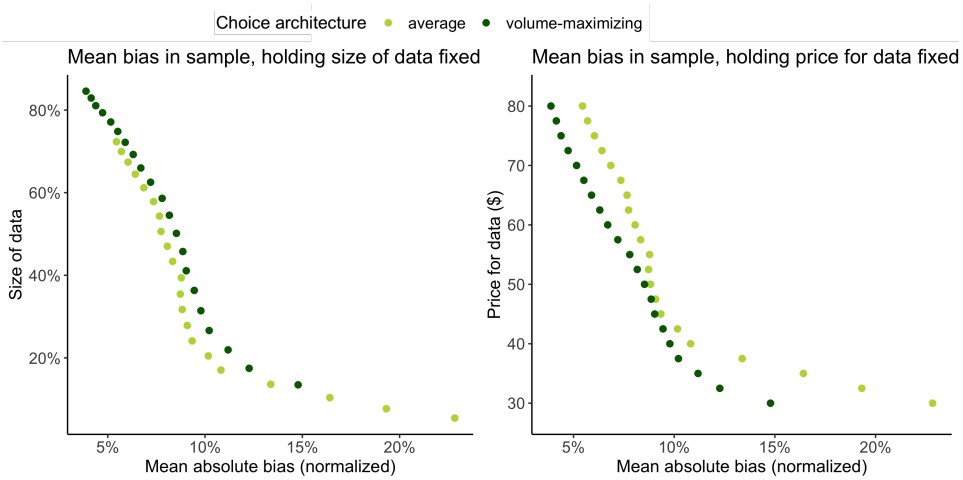


Fig. 3: Data supply curves across choice architecture treatments

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

4.4 The role of choice architecture personalization

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

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

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

5. CONCLUSION

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

The welfare impact of recommendation algorithms

LAURA DOVAL
Columbia Business School and CEPR
and
ALEX SMOLIN
Toulouse School of Economics and CEPR

In this letter, we summarize our recent work on the welfare impact of recommendation algorithms and propose questions for further study. We model recommendation algorithms as an information structure, which shapes how a third party takes actions that affect the welfare of different individuals in a population. Each recommendation algorithm thus induces a welfare profile, describing the expected payoffs of different individuals when the third party takes actions following the algorithm. Our framework allows us to characterize and compute the set of all such profiles, which we dub the Bayes welfare set. The Bayes welfare set allows us to reduce society's choice of an algorithm to the choice of a Bayes welfare profile. Our framework complements that of the algorithmic fairness literature which remains agnostic about the population's payoffs, focusing instead on statistical properties of algorithms, such as accuracy, parity, or fairness.

Categories and Subject Descriptors: [**Social and Behavioral Sciences**]: Economics—

General Terms: Economics, Theory
Additional Key Words and Phrases: Recommendation algorithms, fairness, persuasion, information structures

Authors' addresses: laura.doval@columbia.edu, alexey.v.smolin@gmail.com

1. INTRODUCTION

Information has increasingly become a tool for shaping society's choices in high-stakes domains. While this phenomenon is not exclusive to the "big data" economy—consider the role of scores and ratings in school placement, promotions, and credit allocation—the rise of algorithmic recommendations has highlighted society's growing reliance on information in policy-relevant domains. Consider, for instance, the role of algorithms in recommending who obtains bail [Angwin et al. 2016], credit [Jagtiani and Lemieux 2019], who is hired [Raghavan et al. 2020; Li et al. 2020], and which health treatments to prescribe [Obermeyer et al. 2019], and more recently, generative AI models, which users can leverage as *virtual consultants* [Immorlica et al. 2024].

The ever-increasing role of recommendation algorithms in high-stakes domains and their obvious welfare impact has caught the attention of the Computer Science and Economics communities. For instance, [Kearns and Roth 2019; 2020] underscore the importance of having portable definitions of privacy or fairness, which can be coupled with the training model's objective, to produce algorithms with desirable outcomes. Whereas the literature on differential privacy and algorithmic fairness is agnostic about how to measure individuals well-being or the objective of the algorithm designer, [Mullainathan 2018; Kleinberg et al. 2018; Rambachan et al. 2020] argue for letting the social planner's objective determine the properties of algorithms.

In [Doval and Smolin 2021; 2024], we provide a framework to study the welfare impact of recommendation algorithms on a population of heterogeneous individuals. Our framework marries welfare economics and information design. It integrates welfare economics because a primitive of our environment is a measure of individual welfare, which could represent the actual utility function of individuals in the society, or the social planner's perception of this utility. It also draws from information design because recommendation algorithms fundamentally operate as information structures, which provide noisy signals about an underlying state of the world to a decision maker who ultimately takes actions on behalf of the individuals. As such, recommendation algorithms are inherently bounded in their ability to generate welfare, the same way an information designer is bounded in their ability to persuade a receiver to take a given action [Dughmi 2017].

Our primary goal in this note is to introduce the readers to our framework, based on illustrations of the results in [Doval and Smolin 2021] and [Doval and Smolin 2024]. Section 2 introduces the simplest version of our framework to lay down the concepts in the simplest terms. In Section 3, we extend the framework so that it is closest to that in algorithmic fairness. We conclude by pointing out applications of our framework to information design and directions for future research at the intersection of Computer Science and Economics.

2. BASIC FRAMEWORK

In the basic model, a unit mass population of individuals have types in a finite set $\Theta = \{\theta_1, \ldots, \theta_N\}$, drawn from a full support prior distribution, $\mu_0 \in \Delta(\Theta)$. Each individual's welfare depends on her type θ and an (unmodeled) outside observer's

belief about her type. We represent this by a welfare function $w : \Delta(\Theta) \times \Theta \to \mathbb{R}$, representing for each belief μ and type θ, the welfare of individuals of type θ when the outside observer's belief is μ. For instance, if individuals' welfare depends on the actions of the outside observer, the welfare function captures in reduced-form how the outside observer's action, and hence welfare, changes as the outside observer's beliefs about Θ changes. Alternatively, the welfare function may capture that the population's welfare may be driven by image or reputation concerns, like in [Bénabou and Tirole 2006], or psychological motives, as in [Lipnowski and Mathevet 2018].

We model algorithms as *information structures*. An information structure $\Pi = (\pi, S)$ consists of a countable set of labels S, and a mapping π, which associates to each type θ a distribution over signals $\pi(\cdot|\theta) \in \Delta(S)$. Let μ_s denote the posterior belief given signal $s \in S$. An information structure induces two kinds of distribution over posterior beliefs $\{\mu_s : s \in S\}$. First, for each θ, the signal distribution $\pi(\cdot|\theta)$ induces a distribution over posterior beliefs conditional on an individual's type being θ. Second, the prior μ_0 and the signal distribution induce an *unconditional* distribution over posterior beliefs. We denote them by $\langle \Pi|\theta \rangle$ and $\langle \Pi \rangle$, respectively.

The welfare function w together with an information structure, Π, defines a welfare profile, $w_\Pi : \Theta \mapsto \mathbb{R}$, as

$$w_\Pi(\theta) = \mathbb{E}_{\langle \Pi|\theta \rangle} [w(\mu, \theta)] = \sum_{s \in S} \pi(s|\theta) w(\mu_s, \theta). \tag{1}$$

We denote such a profile, a Bayes welfare profile, and the set of all Bayes welfare profiles, the Bayes welfare set. Formally, the Bayes welfare set is defined as:

$$\mathrm{W} \equiv \left\{ \mathrm{w} \in \mathbb{R}^N : \exists\, \Pi \text{ s.t. } \mathrm{w}_i = w_\Pi(\theta_i)\ \forall i \in \{1, \ldots, N\} \right\}. \tag{2}$$

From the point of view of welfare economics, the Bayes welfare set admits a classical interpretation: It is the utility possibility set in an economy in which information structures take the role of allocations.

An apparent difficulty when characterizing the Bayes welfare set is that the Bayes welfare profiles depend on the *conditional* distributions over posterior beliefs induced by the information structure (cf. Equation (1)). However, we show any Bayes welfare profile satisfies the following:

$$w_\Pi(\theta) = \mathbb{E}_{\langle \Pi|\theta \rangle} [w(\mu, \theta)] = \mathbb{E}_{\langle \Pi \rangle} \left[\frac{\mu(\theta)}{\mu_0(\theta)} w(\mu, \theta) \right] = \mathbb{E}_{\langle \Pi \rangle} [\hat{w}(\mu, \theta)]. \tag{3}$$

That is, the expectation of w under Π conditional on θ can be expressed as the unconditional expectation of the *truth-adjusted* welfare function, $\hat{w}$, under Π. The truth-adjusted welfare function, $\hat{w}$, is the welfare function w adjusted by the *truth-drift* $\mu(\theta)/\mu_0(\theta)$. For any given posterior belief μ, the likelihood ratio $\mu(\theta)/\mu_0(\theta)$ measures the representation of type θ under μ relative to its ex ante representation under μ_0.

It follows that the Bayes welfare set can be characterized by studying the convex hull of the graph of the *vector-valued* function, $\hat{\mathrm{w}} : \Delta(\Theta) \mapsto \mathbb{R}^N$, where for each $i \in \{1, \ldots, N\}$, $\hat{\mathrm{w}}_i(\mu) \equiv \hat{w}(\mu, \theta_i)$. Indeed, we have the following:

Theorem 2.1 [Doval and Smolin 2024, Theorem 1]. *The Bayes welfare set* W *satisfies the following:*

$$\mathrm{W} = \left\{ \mathrm{w} \in \mathbb{R}^N : (\mu_0, \mathrm{w}) \in \mathrm{co}\left(\text{graph } \hat{\mathrm{w}}\right) \right\}, \tag{4}$$

where co *denotes the convex hull operator.*

Theorem 2.1 provides a geometric characterization of the set W: it is the section at the prior of the convex hull of the graph of the truth-adjusted welfare function $\hat{\mathrm{w}}$. We illustrate Theorem 2.1 through an example:

Example 2.2 *Online marketplace.* An online marketplace wants to design a recommendation algorithm, directing consumers to buy from sellers in the platform. For simplicity, assume sellers may be of one of two equally likely types: low quality θ_1, and high quality θ_2. Consumers prefer to buy from high quality sellers. Thus, each seller's profit in the marketplace depends on the likelihood μ the consumer attaches to the seller being of high quality. In particular, we assume the sellers' profits as a function of consumers' beliefs are as follows:

$$w(\mu, \theta) = \begin{cases} 0 & \text{if } \mu \in [0, 1/3) \\ 1/2 & \text{if } \mu \in [1/3, 2/3) \\ 1 & \text{if } \mu \in [2/3, 1] \end{cases}. \tag{5}$$

In this example, the set W then represents the set of profit profiles sellers with different qualities can attain in the marketplace under some information structure.

Figure 1 illustrates the convex hull of the graph of $\hat{\mathrm{w}}$ (Figure 1a) and the Bayes welfare set (Figure 1b) for the online marketplace example. For instance, fully revealing or concealing the sellers' quality is always feasible, so that the full and no-disclosure profiles, w^{FD} and w^{ND} are feasible. We highlight some properties of the Bayes welfare set:

—Despite the welfare function being symmetric across seller types, the Bayes welfare set is not symmetric because the adjusted-welfare function is not symmetric. By Bayes rule, when consumers are optimistic about the seller's quality being high, it is more likely they are facing a high rather than a low quality seller.

—In particular, the Bayes welfare set lies above the 45° line: the only Bayes welfare profile equalizing seller profits is the no disclosure one, but it is not Pareto efficient. In other words, fairness—measured by welfare parity—may be at odds with Pareto efficiency.

—The Pareto frontier of the Bayes welfare set is given by its north-east boundary. In particular, the flat segment at the top shows the profits of low-quality sellers can be increased without decreasing those of high-quality sellers.

—The points on the decreasing part of the Pareto frontier can only be generated with at least three signals. By contrast, in standard Bayesian persuasion, two signals are enough in the case of two states. Formally, the analogue of the W in Bayesian persuasion has dimension N, whereas the W has dimension $2N - 1$.

Because the Bayes welfare set is convex, it can be alternatively described by its supporting hyperplanes. [Doval and Smolin 2024, Theorem 2] shows the frontier of

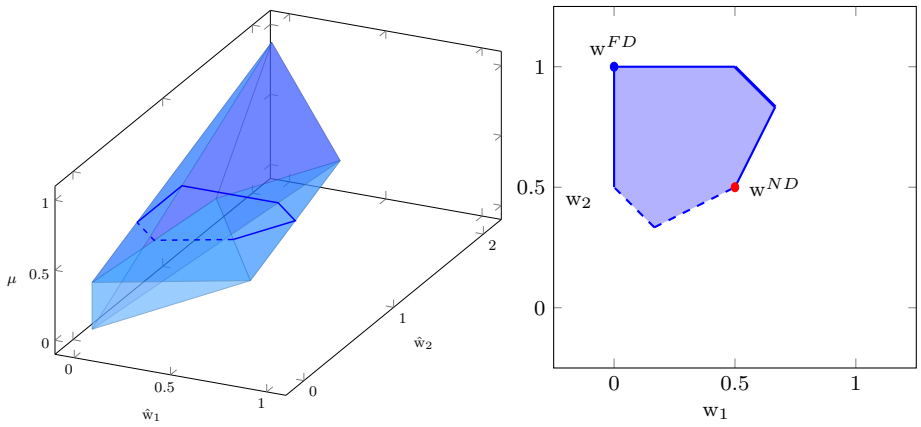


Figure (a) The convex hull of the graph of $\hat{w}$. Figure (b) The Bayes welfare set W.

Fig. 1: Constructing the Bayes welfare set in Example 2.2; w^{FD} and w^{ND} denote the profit profiles under full and no information, respectively.

the Bayes welfare set can be obtained as the solution to series of Bayesian persuasion problems as in [Kamenica and Gentzkow 2011], in which a utilitarian planner takes the role of the information designer. Concretely, consider the supporting hyperplane of the W in direction $\lambda \in \mathbb{R}^N \setminus \{0\}$. Then, the Bayes welfare profiles on the boundary of the W in direction λ can be obtained by solving the Bayesian persuasion problem of a sender with indirect utility

$$\hat{v}_\lambda(\mu) = \sum_{\theta \in \Theta} \mu(\theta) \frac{\lambda(\theta)}{\mu_0(\theta)} w(\mu, \theta).$$

We have found this result very useful in computing the Bayes welfare set in applications (see also [Corrao and Dai 2023] for an application to strategic communication).

3. BEYOND THE BASIC MODEL: GROUPS AND DATA

Two assumptions are implicit in the analysis so far. First, we assume the variable the unmodeled outside observer cares about is the same variable on which we condition the payoffs. Consider, however, an employer making hiring decisions based on a candidate's ability. If candidates belong to different groups, basing hiring recommendations on ability impacts the welfare of candidates across different groups. Second, we assume the information structure can arbitrarily condition on an individual's payoff-relevant type. However, because regulation may prevent the disclosure of protected characteristics, such as gender or race, considering algorithms that respect these restrictions is natural whenever θ includes such characteristics.

Formally, we extend the basic model as follows. We now distinguish between three random variables: an individual's group $g \in G$, the state $\omega \in \Omega$, and data $d \in D$. The first is the variable we condition payoffs on; the second is the variable of interest to the outside observer; the third allows us to capture limits on the information provided. We let $\mathbb{P} \in \Delta(G \times \Omega \times D)$ denote the joint distribution over group-state-

data pairs, and in a slight abuse of notation we denote by $\mathbb{P}(\cdot|g)$ and $\mathbb{P}(\cdot|d)$ the prior distribution conditional on the individual's group and the data realization, respectively. Below, we denote the marginal of $\mathbb{P}$ on D by $\eta_0 \in \Delta(D)$. In a slight abuse of notation, we define the welfare function as $w : \Delta(\Omega) \times \Omega \times G \mapsto \mathbb{R}$, with its first argument being the (unmodeled) outside observer's belief μ about the state ω, $\mu \in \Delta(\Omega)$. The basic model corresponds to the case in which $\Theta = G = \Omega = D$ and $\mathbb{P}(\omega, d|g) = \mathbb{1}[g = \omega = d]$.

To capture the limits data imposes on information provision, an information structure is now defined as a tuple (π, S), where $\pi : D \mapsto \Delta(S)$. Given the information policy, belief updating about (g, ω, d), and hence about ω, depends only on the updated belief about d. Specifically, letting η_s denote the updated belief starting from η_0, after observing signal $s \in S$, the updated belief on (g, ω, d) is given by $\mathbb{P}(g, \omega|d)\eta_s(d)$. Without loss of generality, we can write the welfare function as $w_\dagger(\eta, \omega, g) \equiv w(\mu(\eta), \omega, g)$.

Given an information structure (π, S), the welfare of individuals of group g is:

$$w_\Pi(g) = \sum_{\eta \in \text{supp}(\Pi)} \sum_{s \in S : \eta_s = \eta} \sum_{(\omega,d)} \mathbb{P}(\omega, d|g)\pi(s|d)w_\dagger(\eta, \omega, g), \tag{6}$$

and the Bayes welfare set continues to be defined as the set of Bayes welfare profiles.

The characterization of the Bayes welfare set in the basic model extends verbatim to the more general model, once we observe (the analogue of) the truth-adjusted welfare function now takes the form:

$$\hat{w}_\dagger(\eta, g) = \sum_{(\omega,d)} \mathbb{P}(\omega, d|g) \frac{\eta(d)}{\eta_0(d)} w_\dagger(\eta, \omega, g). \tag{7}$$

Equation (7) allows us to provide further insight into the adjusted welfare function in the basic model. The likelihood correction is now based on the variable d, highlighting that it corresponds to the variable on which information is being provided. In addition, the presence of additional uncertainty requires averaging over ω and d using weights $\mathbb{P}(\omega, d|g)$. Thus we can immediately extend Theorem 2.1 as:

Theorem 3.1 [Doval and Smolin 2021, Theorem 4]. *The Bayes welfare set can be calculated as:*

$$\mathrm{W} = \left\{ \mathrm{w} \in \mathbb{R}^{|G|} : (\eta_0, \mathrm{w}) \in \mathrm{co}\left(\mathrm{graph}\, \hat{\mathrm{w}}_\dagger\right) \right\}. \tag{8}$$

We note again that it is the prior on data, η_0, and not on the states which determines the constraint on how much information can be provided about the state of the world, and hence the limits on how much welfare can be generated via information.

Example 3.2 *Data Regulation in Hiring.* Consider two equally likely groups of workers, labeled A and B. Workers can have one of two ability levels $\Omega = \{0, 1\}$. In each group, half of the workers are high ability and half are low ability. Suppose these workers face a competitive job market: if the market's perceived likelihood that their ability is 1 equals $\mu \equiv \mu(1)$, they receive wage equal to μ. Equating workers' welfare to their wages, this means that $w(\mu, \omega, g) = \mathbb{E}_\mu[\omega]$.

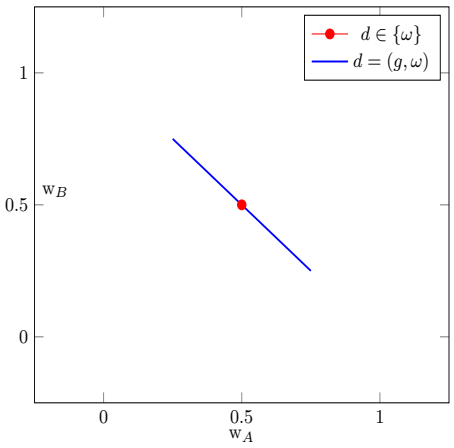


Fig. 2: Bayes welfare set under different data regimes in Example 3.2: The red circle corresponds to the Bayes welfare set under data regimes (i)–(iii); the blue line is the Bayes welfare set in regime (iv).

Rather than assuming a fixed data structure, we compare the Bayes welfare sets in this setting across two data regimes which can be interpreted as different data regulation policies that limit how much information can be revealed about a worker's ability: data reveals ability, but not group (i.e., $D = \Omega$), and data reveals both group and ability (i.e., $D = G \times \Omega$). Figure 2 illustrates the Bayes welfare set in each of these regimes.

Whereas in the first regime we can provide meaningful information about ability, that the distribution of ability is independent across groups together with the martingale property of beliefs implies that on average the posterior belief about the ability remains the same as under no information. It follows that in this case the Bayes welfare set consists of the no disclosure profile, $\mathrm{W} = \{(1/2, 1/2)\}$.

Consider now the second regime and an information structure that pools low-ability workers from group A with high-ability workers from group B and fully reveals all other workers. We can represent this as an information structure with signals $\{B0\}, \{A0, B1\}$, and $\{A1\}$, and induced posterior expectations of 0, $1/2$, and 1, respectively. Because different groups induce these signals with different probabilities, each group's welfare is given by:

$$\mathrm{w}_A = \frac{1}{2}\mathbb{E}[\mu \mid \{A0, B1\}] + \frac{1}{2}\mathbb{E}[\mu \mid \{A1\}] = \frac{1}{2}\frac{1}{2} + \frac{1}{2}1 = \frac{3}{4}, \tag{9}$$

$$\mathrm{w}_B = \frac{1}{2}\mathbb{E}[\mu \mid \{B0\}] + \frac{1}{2}\mathbb{E}[\mu \mid \{A0, B1\}] = \frac{1}{2}0 + \frac{1}{2}\frac{1}{2} = \frac{1}{4}. \tag{10}$$

In fact, this information structure achieves the maximal possible payoff for group A: It never pools workers from group A with the low ability workers of group B, it never pools the high ability workers from group A with workers from group B, and it pools all high ability workers from group B with the workers from group A. As such, the maximal welfare w_A is $3/4$.

We note, however, that the average payoff across groups is the same across all information structures:

$$\frac{1}{2}\mathrm{w}_A + \frac{1}{2}\mathrm{w}_B = \frac{1}{2}\mathbb{E}[\mu \mid g = A] + \frac{1}{2}\mathbb{E}[\mu \mid g = B] = \mathbb{E}[\mu] = \frac{1}{2}. \tag{11}$$

In other words, information merely *redistributes* welfare across the groups. Consequently, the information structure that maximizes the welfare of group A minimizes that of group B.

These observations together with the symmetry of the setting imply that in the fourth regime the Bayes welfare set is given by:

$$\mathrm{W} = \left\{(\mathrm{w}_A, \mathrm{w}_B) \in [1/4, 3/4]^2 : \mathrm{w}_A + \mathrm{w}_B = 1\right\}. \tag{12}$$

4. FINAL REMARKS

We conclude by describing alternative applications of our framework as well as some directions for further research.

4.1 Applications to information design

By interpreting our welfare function as an individual's type-dependent payoff function, the Bayes welfare set is also the object of interest in more standard information design applications. For instance, the types may represent the private information of an informed principal who can commit to an information structure only *after* observing her type, as in [Perez-Richet 2014] and [Koessler and Skreta 2023]. Similar constraints appear in the studies of information design without commitment, as in [Lipnowski and Ravid 2020], [Drakopoulos et al. 2022], and [Corrao and Dai 2023]. Thus, the Bayes welfare set can be viewed as a unifying concept that underlies the incentive constraints the equilibrium information structure must satisfy. As we show in our first working paper version, [Doval and Smolin 2021], our tools also open the door to the study of new problems in this literature such as communication equilibrium payoffs in Bayesian persuasion with transparent motives and Bayesian persuasion with an ambiguity averse sender.[1]

4.2 Further research

We conclude with three (non-exhaustive) suggestions for future research:

It is well-known that various statistical notions of fairness, such as equalized odds and calibration, are incompatible with each other (cf. [Chouldechova 2017; Kleinberg et al. 2016]). Furthermore, this incompatibility remains even when considering relaxations [Pleiss et al. 2017]. Yet, the Bayes welfare set provides another way to visualize the trade-offs among these competing notions. For instance, one could use the Bayes welfare set to understand which group is hurt the most when imposing either calibration or equalized odds. Similarly, one could consider information structures that preserve some form of privacy—e.g., the algorithm recommendations do not reveal information about group membership—and study the Bayes

[1] [Corrao and Dai 2023] fully characterize the set of communication equilibria with transparent motives.

welfare profiles consistent with such restrictions (cf. [Gopalan et al. 2021; Strack and Yang 2024]).

Since the seminal work of [Dughmi and Xu 2016], the computer science literature has made incredible progress in algorithmic Bayesian persuasion (see, e.g., [Babichenko and Barman 2016; Arieli and Babichenko 2019; Banerjee et al. 2024]). Most of this work is concerned with the computational aspects of achieving the sender's preferred payoff, whereas our work focuses on the cross-sectional implications of different information structures for which the sender's average payoff may not be a sufficient statistic.

In many applications, considering constraints on the information structures the planner has access to is natural. The model in Section 3 puts limits on how much information can be provided about the payoff-relevant state. Constraints such as those arising from differential privacy are relevant in many applications and understanding how they shape the choice out of the Bayes welfare set is of interest.

REFERENCES

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica 23*, 77–91.

Arieli, I. and Babichenko, Y. 2019. Private bayesian persuasion. *Journal of Economic Theory 182*, 185–217.

Babichenko, Y. and Barman, S. 2016. Computational aspects of private bayesian persuasion. *arXiv preprint arXiv:1603.01444*.

Banerjee, S., Munagala, K., Shen, Y., and Wang, K. 2024. Fair price discrimination. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2679–2703.

Bénabou, R. and Tirole, J. 2006. Incentives and prosocial behavior. *American Economic Review 96,* 5, 1652–1678.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5,* 2, 153–163.

Corrao, R. and Dai, Y. 2023. The bounds of mediated communication. *arXiv preprint arXiv:2303.06244*.

Doval, L. and Smolin, A. 2021. Information payoffs: An interim perspective. *arXiv preprint arXiv:2109.03061*.

Doval, L. and Smolin, A. 2024. Persuasion and welfare. *Journal of Political Economy 132,* 7, 2451–2487.

Drakopoulos, K., Lo, I., and Mulvany, J. 2022. Blockchain mediated persuasion. *USC Marshall School of Business Research Paper Sponsored by iORB*.

Dughmi, S. 2017. Algorithmic information structure design: a survey. *ACM SIGecom Exchanges 15,* 2, 2–24.

Dughmi, S. and Xu, H. 2016. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 412–425.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. 2021. Omnipredictors. *arXiv preprint arXiv:2109.05389*.

Immorlica, N., Lucier, B., and Slivkins, A. 2024. Generative ai as economic agents. *ACM SIGecom Exchanges 22,* 1, 93–109.

Jagtiani, J. and Lemieux, C. 2019. The roles of alternative data and machine learning in fintech lending: Evidence from the lendingclub consumer platform. *Financial Management 48,* 4, 1009–1029.

Kamenica, E. and Gentzkow, M. 2011. Bayesian persuasion. *American Economic Review 101*, 2590–2615.

Kearns, M. and Roth, A. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kearns, M. and Roth, A. 2020. Ethical algorithm design. *ACM SIGecom Exchanges 18,* 1, 31–36.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*. Vol. 108. 22–27.

Kleinberg, J., Mullainathan, S., and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Koessler, F. and Skreta, V. 2023. Informed information design. *Journal of Political Economy 131,* 11, 3186–3232.

Li, D., Raymond, L. R., and Bergman, P. 2020. Hiring as exploration. *National Bureau of Economic Research*.

Lipnowski, E. and Mathevet, L. 2018. Disclosure to a psychological audience. *American Economic Journal: Microeconomics 10,* 4, 67–93.

Lipnowski, E. and Ravid, D. 2020. Cheap talk with transparent motives. *Econometrica 88,* 4, 1631–1660.

Mullainathan, S. 2018. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 1–1.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366,* 6464, 447–453.

Perez-Richet, E. 2014. Interim bayesian persuasion: First steps. *American Economic Review 104,* 5, 469–74.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. 2017. On fairness and calibration. *Advances in neural information processing systems 30*.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.

Rambachan, A., Kleinberg, J., Ludwig, J., and Mullainathan, S. 2020. An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*. Vol. 110. 91–95.

Strack, P. and Yang, K. H. 2024. Privacy preserving signals. *Available at SSRN 4467608*.

Online Advertisements with LLMs: Opportunities and Challenges

SOHEIL FEIZI, MOHAMMADTAGHI HAJIAGHAYI, KEIVAN REZAEI, SUHO SHIN
University of Maryland

In modern online platforms, advertisement plays a crucial role in subsidizing their operational costs. With the emerging advances in generative AI, this position piece explores the potential of running online advertisement systems for Large Language Models (LLMs). Our position is that, the existing framework for modern search advertisement system does not carry over to LLM advertisement. To understand how one could run advertisement systems for LLMs by integrating ads into their outputs and facilitate future research in this area, we introduce a generic framework for LLM advertisement. We explore several aspects of design choices within our framework and propose potential research questions to operate a practical system.

Categories and Subject Descriptors: J.4 [**Social and Behavior Sciences**]: Economics

General Terms: Algorithms, Design, Economics, Theory

Additional Key Words and Phrases: Online advertising, auction design, large language model

1. INTRODUCTION

In the vast landscape of online search engines, the role of advertisements has become pivotal, shaping the digital experience for users globally. The enormity of the market, with trillions of dollars at play, shows the economic significance of advertising, *e.g.,* the market size of search advertisement as of 2020 was valued at USD 164.12 billion [Insight 2022]. Online advertising not only brings revenue to platform companies but also plays a crucial role in subsidizing free access to information and services for every individual. The democratization of education, information, and other resources is made possible through ad-driven models, enabling individuals to not only consume content for free but also contribute to the digital ecosystem. Additionally, the symbiotic relationship between advertisements and content creation creates a feedback loop, fostering economic growth. As a side note, even subscription-based streaming platforms like Netflix are starting ad-supported plans [Netflix 2022].

On the other hand, recently, large language models (LLMs) [Brown et al. 2020; Anil et al. 2023; Thoppilan et al. 2022] have gained widespread adoption among users, serving various functions including question answering, content generation, translation, code completion, and more [Nijkamp et al. 2022; Fried et al. 2022; Wang

Authors' addresses: (sfeizi,hajiagha,krezaei,suhoshin)@umd.edu
Version history. Nov 2023: first version. Feb/Apr/Sep 2024: added discussions on recent works. This version: revised presentation. Full version is available at [Feizi et al. 2023].
We thank Sébastien Lahaie, Irene Lo, and Sam Taggart for valuable feedback. The authors are partially supported by DARPA QuICC, NSF AF:Small #2218678, NSF AF:Small #2114269, Army-Research Laboratory (ARL) #W911NF2410052, and MURI on Algorithms, Learning and Game Theory.

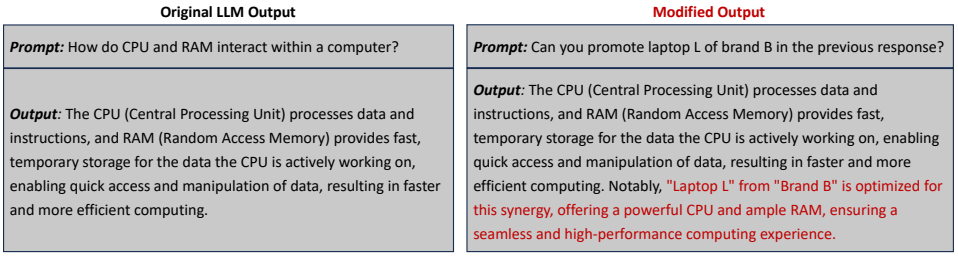


Fig. 1: An example of providing **unstructured** advertisement in LLM output. The left figure shows the original prompt to and output from ChatGPT 4. The right figure shows the prompt used to **incorporate an ad**.

and Komatsuzaki 2021; Liu et al. 2023]. The proliferation of AI-driven assistant language models, such as ChatGPT, has contributed to a growing trend wherein individuals increasingly use such models to address their inquiries, occasionally replacing traditional search engines as their primary information-seeking tool. According to [PCMag 2023], even for now, 35% of casual users say they find LLMs to be more helpful in finding information than search engines. It is obvious that such a trend will be accelerating in the near future as well. The substantial usage volumes stemming from diverse users would induce companies offering these tools, which we call *LLM providers*, to contemplate revenue generation through advertising [AdWeek 2023; Crunch 2023; Microsoft 2023]. Consequently, an interesting and fundamental question arises:

How can **LLM providers** make revenue by running an online **advertisement** on their services?

The concept of online advertising has been extensively studied within the realm of search engines, where auctions are conducted among advertisements from advertisers when a user inputs a query. This paper focuses on the prospect of transposing this online advertising model and auction framework to the context of large language models. We further discuss technical challenges and potential framework to run online advertisement system in LLM, thereby calling academic and industrial researchers to the area of importance.

Search advertising. To better explain fundamental differences between standard search advertising (SA)[1] and LLM advertising (LLMA), we briefly introduce how standard SA works [Lahaie et al. 2007]. (1) *Bidding*: In SA, the owner of each ad i writes bid $b_i \in \mathbb{R}_{\geq 0}$ on targeting *keyword* for $i \in [n]$, which can be a set of keywords. (2) *Output generation*: The platform first decides *how many slots* to allocate for ads in the search engine results page (SERP), say k. (3) *Prediction*: Given k slots in SERP, the platform then predicts the click-through-rate (CTR) α_{ij} when ad i is

[1]Its mechanism design problem is often called sponsored search auction (SSA).

allocated in slot j. (4) *Auction*: The platform then optimizes

$$\max_{x\in[0,1]^{n\times k}} \sum_{i=1}^{n}\sum_{j=1}^{k} \alpha_{ij} b_i x_{ij}, \tag{1}$$

where $x = (x_{ij})_{i\in[n],j\in[k]}$ is the (possibly randomized) allocation vector such that $x_{ij} = 1$ if ad i is allocated in slot j given the constraint $\sum_{i=1}^{n} x_{ij} \leq 1$ for every $j \in [k], i \in [n]$. The platform then charges each ad according to some pre-defined payment rule.

Overall, whenever a user arrives in the platform and searches a keyword, the set of ads related to the keyword are determined. Then, the platform collects corresponding bids of the selected ads, decides the number of slots k, predicts CTR, and the auction runs.

Motivating example. How would LLMA be fundamentally different from SA? We start with illustrative scenarios where a user asks a technical question about computers (Figure 1). Without advertisement, an LLM would typically generate a response to address the user's query. To incorporate advertisements in the generated output, there is a spectrum of possibilities for including ad content, such as: (a) putting the ads outside the response but visibly in the user interface, (b) incorporating the ads within the generated output directly. The ads in option (a) can be treated as display ads, and may be relatively easy to handle using some of the vast amount of prior work on display ads. However, (b) is more similar to a sponsored search ad.[2] We will focus on approach (b), which will entail fundamental challenges that have not arisen in traditional SA. [3]

SA versus LLMA. Recall the process of bidding, output generation, prediction, and auction in the SA mentioned before, and imagine implementing those modules for LLMA.

For the bidding module, since the LLM's query could be far much complicated than just a single keyword in SA, how could the advertisers express their willingness-to-pay for each query? Essentially, each advertiser might want to significantly adjust its bid based on how much they find the query to be relevant to its ad. Further, given that the marketing impact will significantly depend on how the LLM incorporates the ad in the output, it is not even clear what is the advertiser's *value* for being included in the ad. If the advertiser's value depends on the generated output, how could the advertiser reflect their *willingness-to-pay* with respect to the output, which might not be accessible in advance? Even further, how can we generate output that smoothly incorporates the ad without hurting the user experience while satisfying the advertiser?

For the prediction module, most SA run an online learning algorithm to update the ad's feature vector with respect to user context [McMahan et al. 2013]. This was

[2]Search ads usually capture user attention better than display ads, *e.g.,* almost 50% more views [Outbrain 2023].

[3]If ads are included within the generated output, there are possibilities of structured outputs (e.g., ads replacing one of the elements in a given list of elements), or the ads can be included beyond the output (similar to display ads). We focus on unstructured output, as it is broadly applicable, while structured output may be addressed using the standard SA framework.

possible because the ad images, hyperlinks, and more generally how they appear in the SERP remain the same across many user interactions. LLM, however, could incorporate ads in a very different manner for each query, which makes it difficult for the LLMA to *learn the CTR*. Also, since the ads are merged into the generated output, they significantly affect user experience. How, then, can we guarantee and measure user satisfaction?

Finally, which kind of auction format should the LLMA run? How can the LLMA adapt for advertising multiple ads in a single output? What would be a reasonable analogue of the autobidding system prevalent in the modern online ad system?

All these questions are not straightforward to answer and to our knowledge have not been formally discussed in the literature.

Outline. Given the outlined uniqueness and differences with SA, we expect that LLMA requires a number of research questions to operate in practice. To understand LLMA and present its technical challenges compared to the SA, we first introduce a generic framework to operate LLMA.[4] Similar to SA, our framework consists of four modules, though the implementation of each module will be very different from SA: (i) *modification* in which the original output of LLM is modified; (ii) *bidding* that advertisers utilize to bid on the modified outputs; (iii) *prediction* in which LLMA computes required information about advertisements; and (iv) *auction* in which the advertisers compete and the final output is selected. We introduce design choices for each module, evaluated against criteria essential for a sustainable system, and discuss the research challenges inherent to each module. Our framework further enables a unified interpretation and comparison of recent approaches for LLM advertisement systems.[5]

1.1 Related works

Here, we discuss related works on LLMs, online ads, and their intersection.

Large Language Models. Advancements in AI, NLP, and conversational agents, driven by Transformer architecture [Vaswani et al. 2017], have given rise to models like GPT-3 [Brown et al. 2020] and BERT [Devlin et al. 2019]. These models revolutionize chatbots, enabling context-aware, human-like interactions across diverse domains [Abd-Alrazaq et al. 2020; Nicolescu and Tudorache 2022]. Everyday use of language models has led researchers to investigate the content generated by these models to ensure that they do not hallucinate [Guerreiro et al. 2023; Ji et al. 2023; Li et al. 2023; Zhang et al. 2023] in their outputs, and do not generate harmful or biased content [Liang et al. 2021; Navigli et al. 2023; Kirk et al. 2021; Shen et al. 2023; Weidinger et al. 2021; Liu et al. 2023]. In fact, trustworthy of LLMs is actively studied by researchers [Liu et al. 2024].

Online Advertisement. Online advertising, particularly within the context of sponsored search auctions, has evolved in recent years, with notable contributions from prior research. Sponsored search auctions have been a subject of extensive

[4]While our framework could serve as an initial step for future research, our focus is to address key questions for the practical operation of LLM advertisement

[5]We briefly discuss further perspectives on the interplay between LLMs and online advertising systems, *e.g.,* improving the user attraction by personalizing the ad images by LLMs, see Section 5 or the full paper [Feizi et al. 2023].

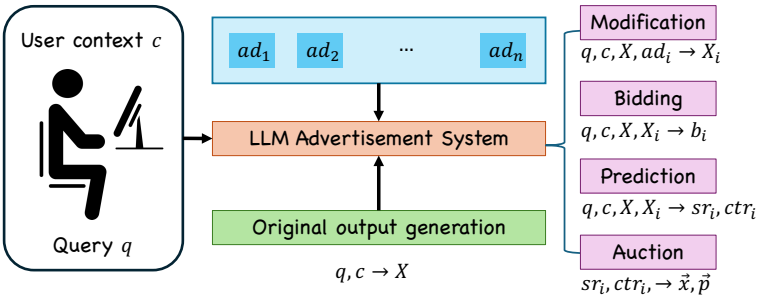


Fig. 2: Overall framework of LLMA.

investigation, emphasizing the optimization of bidding strategies and keyword relevance. [Edelman et al. 2007] provided valuable insights into the economics of sponsored search auctions, shedding light on the dynamics of keyword auctions. [Goel et al. 2009] proposed a contract auction between the advertiser and the publisher, and introduce impression-plus-click pricing for sponsored search auction as an application. We refer to the book by [Roughgarden 2010] for more details. Finally, there exists an emerging interest from the community to study ad auctions in LLMs, which we elaborate more on Section 4.5.

2. FRAMEWORK FOR LLMA

To understand how LLMA could operate in practice, given the stark differences between SA and LLMA, we here present a potential generic framework for LLMA.[6] Mainly, we focus on a scenario in which a user provides a query q to the LLM, andLet the original output by the LLM is given by X. Further, a context c captures a variety of features that are relevant to the advertisement recommendation, *e.g.,* history of the previous queries, user segment, region, and date.[7] Although the number of advertisers varies from time to time, when the user inputs the query q, we suppose that there are n advertisers (bidders) indexed by $\mathrm{adv}_1, \ldots, \mathrm{adv}_n$, each of which is equipped with a single advertisement (ad) ad_i he/she wants to post.

Overall, we divide the LLMA into 4 modules as follows based on their functionalities[8]: (i) output modification, (ii) bidding, (iii) prediction, and (iv) auction. This overall framework is illustrated in Figure 2. In short,

(1) The user writes query q possibly with context c.

(2) The LLMA generates the original output X.

(3) The **modification module** creates modified output X_i per ad.

(4) The **bidding module** generates corresponding bid b_i.

[6]While entirely different systems outside this framework may emerge, we believe it can foster future discussions to address technical challenges in LLMA.

[7]Note here that we assume LLMA could collect such information from the user similar to the standard search engines, but our model accommodates the setting without such data.

[8]These functionalities may be distributed to multiple market players such as supply side platform, display side platform, or ad exchange, as does in the current online ad eco-system.

(5) The **prediction module** predicts user satisfaction rate (SR) sr_i and click-through-rate (CTR) ctr_i.
(6) The **auction module** determines the final output and corresponding payment to charge the selected advertiser.[9]

For the rest of the section, we will explain each module and its responsibility/functionality in a sequential manner.

2.1 Modification module

The modification module generates modified output based on the ads' textual information. This module takes the pair of (q, X, c) and a set of advertisements as the input, and returns $(X_i)_{i\in[n]}$ where X_i denotes the modified output for ad_i.

Overall, we consider two design choices: [10]

(1) In the *advertiser modification* model, the role of generating the modified output is delegated to each advertiser.
(2) In the *LLMA modification* model, LLMA directly generates the modified output.

In Section 4.1, we explore ways to reduce the computational load of generating candidate outputs.

Comparison to SA. Note that the standard SA does not have a modification module explicitly as it is trivial to incorporate ads in each slot. Thus, this is a unique challenge appearing in LLMA. Further, we note that LLM is only directly used in the modification module as well as in generating the original output X, whereas other modules do not require LLM to operate them. Nevertheless, the other modules have technical challenges specific to LLMA, which we now present.

2.2 Bidding module

The bidding module generates bids based on modified outputs. It takes the query q, context c, and modified outputs $(X_i)_{i\in[n]}$ as input and outputs bids $(b_i)_{i\in[n]}$, representing each advertiser's private valuation of impressions, clicks, or conversions. We consider two design choices.

(1) In the *dynamic bidding* model, we provide the query q, context c, original output X,[11] and modified output X_i to the bidder for each query, who then returns the bid.
(2) In the *static bidding* model, each bid is based on keywords from a pre-committed contract, without further communication with the advertiser.

The static bidding model operates by extracting keywords from q, determining relevant ads, and requiring advertisers to set targeted keywords. In Section 4.2, we discuss extensions of these bidding models that allow more flexible bidding strategies. This approach can be preferred when the bid unit accurately represents the

[9] We first focus on presenting a single advertisement in the LLM output, however, generalization of the proposed framework to incorporate multiple ads at once is discussed in Section 4.4.

[10] For either case, one can generate the modified output by giving an additional query to the LLM as presented in Figure 1.

[11] An encrypted context $\hat{c}$ or no context could be considered for privacy.

advertiser's true valuation, regardless of output quality, such as clicks or conversions.

Dynamic bidding models interest advertisers with their measures to estimate the quality of modified output. For instance, if LLMA only considers CTR and ignores user experience, advertisers may wish to adjust bids based on their output quality assessment. Despite a high likelihood of clicks or conversions, poor user experience with low-quality outputs could harm satisfaction with the advertised product.

Comparison to SA. In SA, since the context of the user query is more explicitly represented as keywords, advertisers can safely bid on each relevant keyword (which SA requires). In LLMA, however, it might be difficult to extract proper keywords from the query as the query itself tends to be much longer than in traditional SA due to the flexibility in LLM's query. On the other hand, one can instruct the LLM to extract core keywords or use word embeddings to calculate the similarity between the modified output and query. Further, the generated output significantly affects the marketing impact of the ads in LLMA, whereas it is typically independent from other ads / contents shown within SERP in SA. Finally, the dynamic bidding model exhibits unique challenges of dynamically adjusting the bid with respect to the output, which could further be delegated to another market player or LLM agent.

2.3 Prediction Module

The prediction module computes the user's satisfaction rate (SR) and click-through rate (CTR). The SR measures user satisfaction with the output and influences the decision-making process of the language model to avoid disappointing outputs. The CTR represents the likelihood of the user clicking on the ad link in the output and is crucial for determining auction winners, as it directly impacts the LLMA's revenue. Specifically, if an advertiser's bidding method is cost-per-click (CPC), the expected revenue from the ad is calculated as CPC multiplied by the CTR. Overall, both SR and CTR are functions of the original output X, modified output X', query q, and context c, which returning a real value in $[0, 1]$.

Comparison to SA. Different from the traditional SA whose output is static (ad image/hyperlinks), LLMA constructs textual outputs in a complicated manner. This makes it more difficult for the prediction module to learn the CTR. Moreover, in LLMA, the prediction of user satisfaction is much more directly affected by the incorporation of the ads in the output. This is in stark contrast with SA, where user experience is usually affected only by the number of ad slots/positions in SERP. Detailed methodologies for estimating/learning these functions will be discussed in Section 4.3.

2.4 Auction module

Having computed all the required parameters, we run the auction module to determine the auction winner and the advertiser's charge. The input to the auction module is the set of tuples $(\text{bid}_i, \text{sr}_i, \text{ctr}_i)$ $i \in [n]$, representing bid amount, satisfaction rate, and click-through rate for each bidder. The auction module outputs an (possibly randomized) allocation $\vec{x} \in [0, 1]^n$ and payments $\vec{p} \in \mathbb{R}^n_{\geq 0}$. Specifically, the module determines the auction format, including the allocation function, which

selects the ad, and the payment function, calculating the advertiser's payment to LLMA.[12]

The main goal of LLMA is to maximize its long-term revenue by balancing short-term revenue with user retention. The objective is modeled as a function from bid amount, CTR, and SR to a nonnegative score for a modified output, i.e., selecting $i^* = \text{argmax}_{i \in [n]} \text{Obj}(\text{sr}_i, \text{ctr}_i, \text{bid}_i)$. We do not detail the objective function choice, given the extensive literature on sponsored search auctions. After designing the score function, an auction format should be determined, with many mechanisms available, such as VCG auction or generalized second price auction as desired.

Comparison to SA. The main difference with SA is, similar to what is discussed in the previous subsection, that user satisfaction is a much more important measure to account for. For example, the typical objective in SA (see (1) in Section 1) is social welfare which only accounts for platform and advertisers' utility. In LLMA, however, one might also need to consider user's utility as a function of predicted SR, which would change the allocation function of the mechanism and the payment correspondingly.

3. MARKET PLAYERS AND DESIDERATA

Given the potential framework for LLMA and the presented design choices, we outline several crucial aspects for evaluating each design choice's feasibility and practicality.

In modern search or display ad auctions, numerous market players are involved, including advertisers, users, and platforms. The platform itself is often divided among several players like the demand-side platform, supply-side platform, publisher, and ad exchange. In essence, the ad platform must balance these players' utilities to create a sustainable ecosystem. For example, if the platform shows too many ads in response to a user's query, user retention will likely decrease, discouraging advertisers from using the platform, and eventually harming the ecosystem. Similarly, for LLMA to sustain long-term revenue growth, it must balance the utilities of market players. We will outline the key aspects of the most crucial players: the user, the advertiser, and the platform.

3.1 Player's incentive

User experience. When adding advertisements to LLM output, maintaining high content quality is crucial. Users dislike excessive or irrelevant ads, which can degrade the output and reduce user satisfaction and retention. In modern online ad systems, floor prices are used to filter out irrelevant ads and preserve user experience quality. Similarly, for LLM services, we must ensure that the final output, including ads, remains a high-quality response and closely aligns with what the LLM would originally generate.

Advertiser experience. Advertisers pay the LLMA to include ads in outputs, expecting their products or services to be showcased compellingly. Ads should be engaging and interesting to users, efficiently driving revenue for the advertisers at

[12]This form doesn't allow for adjusting the final output to balance multiple advertisers' preferences, unlike auctions that consider bids and preferences, as in [Duetting et al. 2023].

smaller costs. It is worth noting that advertisements may potentially reduce the overall number of users engaging with the system, which could have adverse effects on the LLMA itself.

Platform revenue. Revenue is LLMA's primary goal, so it must ensure that the additional cost of advertisements is covered by the revenue from advertisers. This is especially critical for LLMA compared to SA, due to the higher computational costs and infrastructure required to run LLMs.

3.2 Desiderata

To balance the utilities of all players and ensure a sustainable LLMA, the following criteria should be considered from the platform's perspective.

Output quality. The LLM's textual output should align with the user's preferences and the query. This involves (a) ensuring the ad is relevant to the user's query and (b) making sure the LLM output aligns with the user's preferences. This includes standard LLM evaluation criteria like accuracy, relevance, coherence, and comprehensiveness.

Additionally, the output should reflect advertisers' preferences, ensuring their ads are integrated as desired and their bids accurately represent their preferences for the query and modified output.

Allocational objective and revenue. As a mediator, LLMA can optimize social welfare to achieve allocational efficiency, i.e., allocate ads to maximize social welfare deterministically or try to maximize its revenue. On the other hand, retrieval-augmented generation (RAG) by [Lewis et al. 2020] that probabilistically retrieves a relevant document from a database of factual documents, is shown to enhance LLM performance by managing ambiguity and factuality, diversifying output, and improving robustness to noise and errors one might want to randomly allocate the ad rather than deterministically to improve the quality of the output. Therefore, LLMA's allocation objective should balance these aspects while ensuring sufficient revenue to avoid operating deficits.

Latency. In LLMA, users expect rapid interactions, similar to search auctions, where prompt responses are typical. Adding advertisements to LLM output introduces some latency, but this should be minimal to avoid disrupting the user experience. The latency requirement for LLMA could be less strict than for search auctions because LLMA generates output word-by-word, whereas search auctions need to retrieve all ads and results immediately upon a query.

Reliability and privacy. LLMA must also address potential risks from advertisers, ensuring system reliability and alignment by considering all possible adversarial behaviors [Hendrycks et al. 2020], *e.g.,* harmful contents or spam links in ads. Additionally, maintaining user privacy is crucial. All user context, information, and data must be kept secure (or encoded) to prevent privacy risks from inadvertent disclosure.

4. CHALLENGES

Recall that our overall framework consists of four modules: modification, bidding, prediction, and auction modules. For each module, we address characteristics,

technical challenges, and research questions relevant to practical implementation and evaluation based on the criteria defined in Section 3.

4.1 Modification module

Challenge: **An advertiser modification model should ensure alignment with user preferences and satisfaction while addressing privacy, reliability, and latency issues.** In the advertiser modification model, LLMA must provide q, X, and C to each advertiser adv_i, which can lead to privacy issues by disclosing user information. Addressing this involves partial or indirect information disclosure, possibly using encryption or differential privacy to protect user data while ensuring high-quality outputs.

Additionally, the model faces reliability issues as advertiser-modified outputs might include illegal or spam content, which may degrade user satisfaction compared to original LLM outputs. LLMA may require an additional module to ensure robustness against such adversarial behavior, but this adds cost of computational resources and latency.

Moreover, increased communication between LLMA and advertisers can raise latency. Therefore, developing efficient protocols is crucial for managing a functional online ad system. The advertiser modification model thus requires novel solutions to address privacy, reliability, and latency concerns.

Challenge: **Effectively reflecting advertiser preferences in the output for LLMA modification model.** The LLMA modification model generally faces fewer privacy, reliability, and latency issues compared to others and focuses on enhancing user experience. However, it may not fully align with advertiser preferences in the output modification process, potentially reducing advertiser satisfaction. Thus, it is important to explore methods for better incorporating advertisers' preferences into the modified output, ensuring that it meets their expectations while still improving the user experience.

Prospect: **Balancing the trade-off between LLMA and advertiser modification models.** To improve advertiser satisfaction in the LLMA modification model, one approach is to let advertisers submit indirect indicators of their preferences. Specifically, after receiving the query q, the original output X, and possibly context c, the advertiser provides a document Y (or a list of it) reflecting its preferences. LLMA can then use Y as a prompt to generate the modified output as per RAG framework, allowing more flexibility in capturing advertiser preferences. However, this method introduces additional communication costs, potentially increasing latency.

Challenge: **Reducing the computational burden of generating every potential candidate output.** Both the advertiser and LLMA modification models require generating all possible output candidates for each potential ad. In the LLMA modification model, this involves multiple runs of the LLM's token sequence generation for each ad in the auction. Given the high cost of LLM operations, this approach may not scale well as the number of ads on the platform increases.

Prospect: **Ex-ante allocation without generating potential candidate out-**

puts. One feasible approach is to implement a prefiltering process to reduce the number of ads competing by filtering out less relevant ads. Alternatively, the auction can be run without generating every possible candidate output by using a modular component to predict the characteristics of the modified output for each candidate ad.

For instance, features can be extracted from the query and ad text, and semantic distances can be computed based on text similarity. Advertisers should be aware of this indirect measure and bid accordingly, assuming the semantic distance will represent the expected output quality. The mechanism then uses submitted bids and semantic distances to determine the ad and generate the final output, requiring only a single generation for the allocated ad. Note that [Hajiaghayi et al. 2024] takes this approach.

4.2 Bidding module

Challenge: **Implementing dynamic bidding model without privacy and latency issue.** The main advantage of the dynamic bidding model is its potential to increase advertiser satisfaction, as advertisers can adjust their bids after seeing the modified output. This flexibility can be appealing to advertisers with the technical capability to dynamically set bids. However, the dynamic bidding model may introduce privacy issues if LLMA discloses private information to advertisers. It may also lead to additional latency since the entire set of modified outputs must be delivered to advertisers. Reliability concerns are minimal, as only the bid amount is communicated. In contrast, the static bidding model avoids privacy, latency, and reliability issues but may reduce advertiser satisfaction because advertisers cannot adjust their bids based on the modified output.

Future research could focus on the practicality of dynamic bidding, including developing protocols and algorithms that address privacy and latency concerns. One may also investigate whom the ad market would comprise, when there is a possibility that the advertisers hire a proxy agent to submit bids on behalf of them, and how the proxy agent (or advertisers themselves) can optimize bids in such scenario.

Prospect: Balancing the trade-off between static/dynamic bidding models. In the static bidding model, improving advertiser satisfaction can be achieved by defining a more flexible static function as the contract between the advertiser and LLMA. Specifically, LLMA could propose a contract where bids are determined by an indirect measure of the modified output. For example, LLMA and the advertiser might agree on a contract where the bid is inversely proportional to the similarity distance d between the original output X and the modified output X_i. If X_i significantly deviates from X, the similarity distance will be large, leading to a lower bid from the advertiser due to concerns about user experience.

Advertisers will need to understand how LLMA estimates and defines this distance measure. LLMA could also develop a more refined method for assessing user interest, attention, and relevance for the modified output from the advertiser's perspective, allowing advertisers to choose contracts based on their preferences.

4.3 Prediction module

Challenge: **Efficient and precise implementation of prediction module.** Estimating CTR in LLMA can follow principles similar to those in modern online advertising. We can train a prediction system to estimate $\text{ctr}_i \in [0, 1]$ based on input X_i, q, and c, using user feedback data. For example, factorization machines and online algorithms can be used after feature extraction from user data, context, and queries, as described in [McMahan et al. 2013]. Given the sparse frequency of each X_i in the family of possible outputs, features from q, X_i, and c should be extracted to map to CTR values. User actions like regenerating responses, clicking ads, or exiting the LLM can be used to refine the prediction module.

Challenge: **Relevance/similarity distance measure to estimate user satisfaction.** To estimate SR, one approach is to assume that the original output X is optimal. In this case, the distance between X and the modified output X_i can serve as a measure of SR, since the closer X_i is to X, the higher the expected user satisfaction. For example, one might define the output distance as $d(X, X_i) := ||\Pr(X|q) - \Pr(X_i|q)||$ using a suitable norm. Here, $\Pr(X|q)$ represents the marginal probability of X given q, which can be computed using standard methods from the literature [Vaswani et al. 2017].

More general functions, such as semantic similarity between documents [Mikolov et al. 2013; Cer et al. 2018; Conneau et al. 2019], can be used. One can further implement a calibration layer to ensure it remains well-calibrated with higher accuracy [McMahan et al. 2013]. The key research question is to identify effective measures for predicting user satisfaction when ads are incorporated into the output.

Prospect: **Incorporating distance measures and online learning.** One may consider combining similarity measures and online learning from user feedback for prediction. To learn online SR (or CTR) estimates from user feedback, a useful indicator of whether the user is satisfied with the output is whether the user *regenerates* the output. One may aim to learn a function which outputs the sr, given the query q, modified output X', and context c. This approach does not assume that the original output X is indeed optimal, thereby allowing the possibility that the user may be satisfied with a modified output X_i even though its distance from X is measured is large. This comes at the cost of additional modular component for learning process. One could further consider an online learning model where the similarity distance is also integrated as one of the features for prediction. If the similarity distance has some positive correlations with the true user satisfaction rate, this would increase the accuracy of the prediction. Essentially, an effective way to capture the both advantages of online learning and distance measure should be studied thoroughly.

4.4 Auction module

Challenge: **Incorporating multiple ads in a single output.** Recall that our framework is presented for a setting where a single ad is allocated. One approach to generalize our framework is to repeatedly run the overall procedure and allocate a single ad at once. That is, one can determine an abstract unit that partitions the output, *e.g.,* paragraph, and run our framework for each unit. Another direct approach to extend our framework is, to let the final displayed output X' not

necessarily belong to $\{X_i\}_{i\in[n]}$, but rather interpolates $\{X_i\}_{i\in[n]}$ by prompting LLM to generate output that simultaneously advertises multiple ads. This resembles the approach of aggregating the preference by [Duetting et al. 2023; Soumalias et al. 2024]. By doing so, it might be possible to deliver multiple advertisements in a fair manner, thereby allowing LLMA to bring more revenue by charging multiple advertisers at once.

One subtle issue is that, since each advertiser bids b_i on delivering X_i, they may not want to write the same bid for the balanced output X', thus it may degrade the advertiser's experience. In the static bidding model, as discussed in Section 4.2, this might be handled by committing to a contract based on measures that represents the advertiser's preferences more in a refined manner. In the dynamic bidding model, one approach would be to append an additional step of asking for bids for the final output again to the advertisers.

4.5 Discussion on Recent Approaches

Several recent theoretical approaches have proposed game-theoretic models for ad auctions in LLMs. We discuss how these can be viewed as implementations of our framework, facilitating future research through modular comparisons of components in each approach and highlighting their pros and cons more explicitly.

[Duetting et al. 2023] propose a model where bidders submit bids and distributions over tokens. This aligns with a tokenized version of our model with (i) LLMA modification by aggregating token distributions, (ii) dynamic bidding with adjustments for each query, (iii) no prediction module as it focuses on advertiser perspectives, and (iv) running a token auction. This approach might suffer from issues of latency and reliability, but could possibly maintain high advertiser experience.

[Hajiaghayi et al. 2024] integrate the RAG framework for segment-based ad auctions, where ads are probabilistically retrieved for segments like paragraphs using bids and a notion of relevance. This corresponds to (i) LLMA modification without pre-generating outputs, (ii) static bidding based on clicks, (iii) indirect CTR prediction by measuring ad relevance, and (iv) running a segment auction. Their approach has less privacy, latency and reliability, but advertiser's experience could be largely dependent on how modular components to compute relevance are designed.

[Dubey et al. 2024] introduce the concept of a prominence auction, wherein the allocation function determines both the prominence of each selected advertisement in the output and the specific ads to be allocated. The prominence assigned to each ad influences its representation in the LLM-generated summary, thereby affecting user attention and engagement. This is shown to be generalizing the standard position auction by [Varian 2007]. This aligns with (i) LLMA modification, (ii) static bidding, (iii) no prediction required, as prominence serves as a CTR proxy, and (iv) running a prominence auction.Similar to [Hajiaghayi et al. 2024], the advertiser's experience highly depends on how well the LLM in the modification module constructs the output with desired prominence.

[Soumalias et al. 2024] propose an auction that truthfully aggregates advertiser preferences using reinforcement learning from human feedback (RLHF), widely used in LLMs to align the outputs of LLMs with diverse human preferences. This can

be viewed as (i) LLMA modification, (ii) static bidding, (iii) no prediction needed, focusing only on advertiser perspectives, and (iv) running an RLHF-based auction. This creates greater issues with latency, since every potential candidate output needs to be created and the advertiser's reward function should be communicated accordingly. However, it could better reflect advertiser's preferences.

5. CONCLUDING REMARKS

There are several potential areas beyond those outlined here. As observed in most modern online advertising platforms, the use of autobidders [Aggarwal et al. 2024], which delegate the bidding process to the platform, appears to be a plausible approach. Another promising direction is leveraging LLMs themselves to enhance modern search and display advertising systems by efficiently tailoring ad content to individual users. For instance, LLMs could be integrated into the standard dynamic creative optimization framework, often referred to as responsive advertising [Google 2024]. For more discussions of these perspectives, we refer to the full paper [Feizi et al. 2023].

REFERENCES

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., and Househ, M. 2020. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *Journal of medical Internet research 22,* 7, e16021.

AdWeek. 2023. Microsoft advertisements on AI driven chat based search. `https://www.adweek.com/media/microsoft-details-how-advertising-works-on-bings-ai-driven-chat-based-search`.

Aggarwal, G., Badanidiyuru, A., Balseiro, S. R., Bhawalkar, K., Deng, Y., Feng, Z., Goel, G., Liaw, C., Lu, H., Mahdian, M., et al. 2024. Auto-bidding and auctions in online advertising: A survey. *ACM SIGecom Exchanges 22,* 1, 159–183.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Crunch, T. 2023. That was fast! Microsoft slips ads into AI-powered Bing Chat. `https://techcrunch.com/2023/03/29/that-was-fast-microsoft-slips-ads-into-ai-powered-bing-chat/`.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Dubey, K. A., Feng, Z., Kidambi, R., Mehta, A., and Wang, D. 2024. Auctions with llm summaries. *arXiv preprint arXiv:2404.08126*.

Duetting, P., Mirrokni, V., Leme, R. P., Xu, H., and Zuo, S. 2023. Mechanism design for large language models. *arXiv preprint arXiv:2310.10826*.

Edelman, B., Ostrovsky, M., and Schwarz, M. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review 97,* 1, 242–259.

Feizi, S., Hajiaghayi, M., Rezaei, K., and Shin, S. 2023. Online advertisements with llms: Opportunities and challenges. *arXiv preprint arXiv:2311.07601*.

Fried, D., Aghajanyan, A., Lin, J., Wang, S., Wallace, E., Shi, F., Zhong, R., Yih, W.-t., Zettlemoyer, L., and Lewis, M. 2022. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999*.

Goel, S., Lahaie, S., and Vassilvitskii, S. 2009. Contract auctions for sponsored search. In *International Workshop on Internet and Network Economics*. Springer, 196–207.

Google. 2024. About responsive search ads . `https://https://support.google.com/google-ads/answer/7684791?hl=en`.

Guerreiro, N. M., Alves, D., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. F. T. 2023. Hallucinations in large multilingual translation models.

Hajiaghayi, M., Lahaie, S., Rezaei, K., and Shin, S. 2024. Ad auctions for llms via retrieval augmented generation. *arXiv preprint arXiv:2406.09459*.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Insight, M. 2022. Global Search advertising. `https://www.millioninsights.com/snapshots/search-advertising-market-report`.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys 55,* 12 (Mar.), 1–38.

Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., and Asano, Y. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Vol. 34. Curran Associates, Inc., 2611–2624.

Lahaie, S., Pennock, D. M., Saberi, A., and Vohra, R. V. 2007. Sponsored search auctions. *Algorithmic game theory 1*, 699–716.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems 33*, 9459–9474.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6449–6464.

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds. Proceedings of Machine Learning Research, vol. 139. PMLR, 6565–6576.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, W., Wang, X., Wu, M., Li, T., Lv, C., Ling, Z., Zhu, J., Zhang, C., Zheng, X., and Huang, X. 2023. Aligning large language models with human preferences through representation engineering.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., and Li, H. 2024. Trustworthy llms: a survey and guideline for evaluating large language models' alignment.

McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., et al. 2013. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1222–1230.

Microsoft. 2023. Monetize AI powered chat experiences. `https://about.ads.microsoft.com/en-us/blog/post/may-2023/a-new-solution-to-monetize-ai-powered-chat-experiences`.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Navigli, R., Conia, S., and Ross, B. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality 15,* 2, 1–21.

Netflix. 2022. Netflix Started Ad-supported plan. `https://help.netflix.com/en/node/126831/`.

Nicolescu, L. and Tudorache, M. T. 2022. Human-computer interaction in customer service: the experience with ai chatbots—a systematic literature review. *Electronics 11,* 10, 1579.

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., and Xiong, C. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.

Outbrain. 2023. Native is better than display. `https://www.outbrain.com/blog/native-ads-vs-display-ads/`.

PCMag. 2023. LLM is replacing search engine. `https://www.pcmag.com/news/when-will-chatgpt-replace-search-engines-maybe-sooner-than-you-think`.

Roughgarden, T. 2010. Algorithmic game theory. *Communications of the ACM 53,* 7, 78–86.

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., and Xiong, D. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Soumalias, E., Curry, M. J., and Seuken, S. 2024. Truthful aggregation of llms with an application to online advertising. *arXiv preprint arXiv:2405.05905*.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Varian, H. R. 2007. Position auctions. *international Journal of industrial Organization 25,* 6, 1163–1178.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems 30*.

Wang, B. and Komatsuzaki, A. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. 2021. Ethical and social risks of harm from language models.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Incentive-Aware Machine Learning; Robustness, Fairness, Improvement & Causality

CHARA PODIMATA
MIT

The article explores the emerging domain of incentive-aware machine learning (ML), which focuses on algorithmic decision-making in contexts where individuals can strategically modify their inputs to influence outcomes. It categorizes the research into three perspectives: *robustness*, aiming to design models resilient to "gaming"; *fairness*, analyzing the societal impacts of such systems; and *improvement/causality*, recognizing situations where strategic actions lead to genuine personal or societal improvement. The paper introduces a unified framework encapsulating models for these perspectives, including offline, online, and causal settings, and highlights key challenges such as differentiating between gaming and improvement and addressing heterogeneity among agents. By synthesizing findings from diverse works, we outline theoretical advancements and practical solutions for robust, fair, and causally-informed incentive-aware ML systems.

1. INTRODUCTION

Machine Learning (ML) algorithms are deeply embedded in various aspects of modern life, influencing everything from enhancing daily conveniences and shaping online purchasing behavior to making critical decisions in areas such as hiring, loan approvals, college admissions, and probation rulings. Given the high stakes of these decisions, individuals often have strong incentives to strategically modify the data they provide to these algorithms to secure more favorable outcomes. For instance, individuals might open additional credit accounts or take other steps to improve their credit scores before applying for a loan. In the context of college admissions, applicants may retake standardized tests like the GRE, enroll in test preparation courses, or even switch schools to boost their class rankings, all in efforts to present themselves as more competitive candidates.

Such instances of "strategic adaptation" have been extensively documented across disciplines including Economics, Computer Science, and Public Policy [Björkegren et al. 2020; Dee et al. 2019; Dranove et al. 2003; Greenstone et al. 2022; Gonzalez-Lira and Mobarak 2019; Chang et al. 2024]. The challenge arises when decision-makers deploying ML algorithms fail to account for these adaptations, potentially undermining the original goals of the policies the algorithms are intended to support. For example, in college admissions, a student's decision to change schools solely to improve their class ranking may not necessarily reflect a substantive improvement in their qualifications.

It is important to note that not all strategic adaptations are inherently problematic. Some represent attempts to "game" the system (e.g., switching schools for a better ranking), while others involve genuine efforts at self-improvement (e.g., dedicating more time to study). The distinction between these types of adapta-

podimata@mit.edu

tions underscores the nuanced nature of this phenomenon and its implications for algorithmic decision-making.

What should decision-makers do when individuals are incentivized to alter the data they provide to ML algorithms in pursuit of better outcomes? And even if the learner manages to robustify (or calibrate) their algorithms to account for such behavior, what are the societal implications? These are some of the central questions addressed by the emerging field of *incentive-aware ML* (also known as "strategic classification" or "performative prediction").[1]

The purpose of this article is to provide an introduction to incentive-aware ML and an overview of the key results in the field. We categorize the literature on incentive-aware ML into three main perspectives: *robustness*, *fairness*, and *improvement & causality*. While some papers contain elements of multiple categories, we classify them based on their primary focus or central contribution. Broadly speaking, the "robustness" perspective adopts the viewpoint of the decision-maker, assuming that agents always attempt to "game" the decision rule. The goal in this context is to design algorithms that achieve optimality despite strategic adaptations by the agents. The "fairness" perspective examines the downstream societal impacts of algorithmic decision-making under varying assumptions about the agents' capacity to strategically adapt. Lastly, the "improvement & causality" perspective recognizes that not all strategic adaptations are harmful; in some cases, agents' adaptations in response to decision-making algorithms lead to genuine, fundamental improvements rather than merely fooling the algorithm. The distinctions among these perspectives, as well as the models and settings considered, will be formalized in the following section.

This article is organized as follows: Section 2 formalizes all the different formulations of the incentive-aware ML problem while giving some example reference papers for each modeling assumption; Section 3 presents a breakdown of the main contributions from the literature from the *robustness* perspective; Section 4 outlines the results that have been obtained from the *improvement & causality perspective*; Section 5 discusses the *fairness* perspective. Finally, Section 6 offers some parting thoughts on where the literature stands and where we should go next (according to the author's personal opinions, at least).

2. OVERVIEW OF MODELS

In the problem of incentive-aware learning, there is an interaction between a *principal* (aka *learner*, *decision-maker*) and *agents*. The problem has been studied both in the *offline* (i.e., where there is one decision that is made by the principal and then the interaction stops) and *online* setting (i.e., where there are sequential decisions). Before we outline each setting, let us introduce some common notation.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ the *feature* space and $\mathcal{Y} = \{0, 1\}$ (resp. $\mathcal{Y} \subseteq [0, 1]$ for linear regression) the *label* (resp. response) space. We assume that the label (resp. response) is $y = h^\star(x)$, where $h^\star : \mathcal{X} \to \mathcal{Y}$ is called the *ground truth* function (which is not

[1]Throughout this article, we use the terms "incentive-aware" and "strategic" ML interchangeably. The author prefers "incentive-aware" as it more comprehensively captures the considerations arising from agents' behaviors. However, "strategic" is more commonly used in the literature.

necessarily linear).[2] We will denote by $\mathcal{H}$ the *concept class* where $h^\star$ belongs to.

Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ be the principal's loss function. Different applications of interest (within the general incentive-aware learning literature) call for different loss functions for the principal. Examples of frequently used loss functions for classification tasks include:

(i) $0-1$ loss (e.g., [Chen et al. 2020]): $\ell(y, y') := \mathbb{1}\{\mathsf{sign}(y \cdot y') = 1\}$.[3]

(ii) logistic loss (e.g., [Dong et al. 2018]): $\ell(y, y') := \log\left(1 + e^{-y \cdot y'}\right)$

(iii) hinge loss (e.g., [Dong et al. 2018]): $\ell(y, y') := \max\{0, 1 - y \cdot y'\}$

For the regression tasks, the most commonly used loss is some L_p norm.

As is the case in traditional ML, the choice of loss function for the principal affects what algorithms should be used, and what guarantees can be obtained.

Offline Setting

In the offline setting (e.g., [Hardt et al. 2016]), we assume that the agents' features are drawn from some distribution $\mathcal{D}$. The interaction between the principal and the agent can be viewed as a *Stackelberg game* that plays out as follows:

(1) Nature draws $x \sim \mathcal{D}$.
(2) The principal —without knowing x— commits to (and publicly announces) a decision-making rule $f : \mathcal{X} \to \mathcal{Y}$.
(3) The agents observe f and their point (x, y).
(4) Given f, x, y, the agents choose $\hat{x}(f)$ where $\hat{x}(f; x, y) \in \mathcal{X}$ is the best-response of the agent (given pair (x, y)) to the principal's rule f.
(5) The agent reports point $(\hat{x}(f; x, y,), y)$ to the principal.

In Step (4), we are using $\hat{x}(f; x, y)$ abstractly; we are going to specify how it is computed later on. At a high level, $\hat{x}(f; x, y)$ is such that the agent obtains a better standing with regards to f (e.g., the agent gets classified as $+1$ from f in classification settings); see "Agents' Response" for details. To simplify notation, we write $\hat{x}(f)$ (instead of $\hat{x}(f; x, y)$ when clear from context.

For now, let's assume that when the agents best respond to a decision-making rule, they are *merely* trying to "game" it. We will contrast this approach to the Causality viewpoint, highlighted below.

In the "robustness" perspective, the principal's goal is to find a function $f^\star \in \mathcal{F}$ (where $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ is a hypothesis class over which we are searching) such that:

$$f^\star = \arg\min_{f \in \mathcal{F}} \quad \mathbb{E}_{x \sim \mathcal{D}}\left[\ell\left(h^\star(x), f\left(\hat{x}(f)\right)\right)\right] \tag{1}$$

In words, in the "robustness" perspective for the offline learning setting, the principal's goal is to find a function that minimizes the expected loss between the ground truth label (resp. response variable for regression) and the predicted label (resp. score) that function f assigns to the (potentially) altered datapoint $\hat{x}(f)$.

[2] Some works on incentive-aware linear regression assume that $y = h^\star(x) + \varepsilon$ where ε is some small, zero mean noise, but that will not constitute a material difference in this article.
[3] We use $\mathsf{sign}(x) = 1$ to denote that x is positive and $\mathsf{sign}(x) = -1$ otherwise.

Online Setting

In the online setting (e.g., [Dong et al. 2018; Chen et al. 2020; Ahmadi et al. 2021], the interaction between the principal and the agents happens repeatedly over T rounds. For every round $t \in [T]$, the interaction protocol is the following:

(1) Nature chooses $x_t \in \mathcal{X}$.
(2) The principal (without observing x_t) commits to (and publicly announces) decision-making rule $f_t \in \mathcal{F}$.
(3) The agent observes f_t and their point (x_t, y_t).
(4) The agent chooses $\hat{x}_t(f_t; x_t, y_t)$ such that $\hat{x}_t(f_t; x_t, y_t)$ is the agent's best response (given pair (x_t, y_t)) to rule f_t.
(5) The agent reports point $(\hat{x}_t(f_t; x_t, y_t), y_t)$ to the principal.

As we did for the offline case, for ease of notation, we will simply write $\hat{x}_t(f_t)$ in place of $\hat{x}_t(f_t; x_t, y_t)$ whenever clear from context. In the online setting, we assume that the principal knows the agents' utility function, but not the original point, x_t. The choice of $\hat{x}_t(f_t)$ in Step (4) depends on the agent's utility function; see "Agent's Response" below.

A couple of remarks are in order. First, the sequence $\{x_t\}_{t\in[T]}$ that the nature chooses can be *adversarial*. Second, $\mathcal{F}$ can be a general class of functions. That said, the current literature only focuses on *linear* functions. Third, for the robustness perspective in online learning settings, we again assume that when the agent strategically adapts to a rule f_t, they can not influence their y_t (i.e., y_t remains the same both for x_t and for the misreport $\hat{x}_t$).

When we adopt the robustness perspective, the principal's goal is to minimize *Stackelberg* regret defined as follows:

$$\mathrm{Reg}(T) := \sum_{t\in[T]} \ell\left(h^\star(x_t), f_t(\hat{x}_t(f_t))\right) - \min_{f^{\mathsf{OPT}}\in\mathcal{F}} \sum_{t\in[T]} \ell\left(h^\star(x_t), f^{\mathsf{OPT}}\left(\hat{x}_t\left(f^{\mathsf{OPT}}\right)\right)\right) \quad (2)$$

Note that similar to the offline model, we are comparing the algorithm's performance to the best fixed rule f^{OPT} *had you given the agents the opportunity to best respond*. In other words, we are comparing to the Stackelberg equilibrium rule.

Causality

So far, in both the offline and online settings, we have assumed that even after the agent strategically adapts, their y_t remains the same as it was prior to the adaptation (e.g., when an agent increases the number of credit cards they have, they have not actually improved their creditworthiness; they have merely tried to game the credit scoring system). This meant that *every* strategic adaptation was perceived as "gaming" and hence, the principal was trying to suppress it. However, for some applications of interest (e.g., for school admissions or loan approvals), some types of strategic adaptation are not gaming and should instead be encouraged or incentivized. For example, in a school admissions example, a strategic adaptation that makes the student study more in order to pass the threshold for admission is not gaming; rather, it a way for the student to become a better potential candidate for the school of their choice.

To capture this, some settings in incentive-aware ML assume that in any d-dimensional feature vector, some features are *causal* (i.e., by changing them, the agent can change their actual y) while the rest are *proxy/non-causal* (i.e., by changing them, the agent cannot change their actual y). As a result, agent actions that change causal features have the ability to change the ground truth qualifications of an agent; as such, they can lead to *genuine improvement*, as opposed to the *gaming* which is induced by proxy features. The papers that assume causality of features (e.g., [Miller et al. 2020; Shavit et al. 2020; Bechavod et al. 2021]) use the language of *structural causal graphs* [Pearl 2009] in order to model the causal effects of the agents' different features.

Agents' Response

We next turn our attention to the way in which the agents choose their best response to the principal's algorithm. For an agent with ground truth feature vectors x, we use $u(x, \hat{x}; f)$ to denote the agent's *utility* for reporting $\hat{x}$ when the principal uses classification/regression function f. We focus on utility functions of the form:

$$u(x, \hat{x}; f) := \mathsf{val}(\hat{x}; f) - \mathsf{cost}(x, \hat{x}) \tag{3}$$

where $\mathsf{val}(\hat{x}, y; f)$ corresponds to the *value* that the agent obtains by reporting $\hat{x}$ when the principal uses f, and $\mathsf{cost}(x, \hat{x})$ corresponds to the *cost* that agent incurs for changing their feature from x to $\hat{x}$. There have been two types of value functions that have been primarily used in the literature:

(i) (continuous) $\mathsf{val}(\hat{x}; f) := f(\hat{x})$ (i.e., the value is just the evaluation of the function f for the reported feature $\hat{x}$) (e.g., [Dong et al. 2018; Bechavod et al. 2022; Shavit et al. 2020]).

(ii) (discrete) $\mathsf{val}(\hat{x}; f) := \gamma \cdot \mathbb{1}\{\mathsf{sign}(f(\hat{x})) = 1\}$ (i.e., the agent cares only about being classified as passing a threshold (e.g., [Chen et al. 2020; Ahmadi et al. 2021]). Unless specified otherwise, we will use $\gamma = 1$.

As for the cost function, there have been primarily two families that the literature has considered:

(i) (L_p-norm) $\mathsf{cost}(x, \hat{x}) := \delta \cdot \|x - \hat{x}\|_p$ for some $\delta > 0$ (e.g., [Chen et al. 2020; Ahmadi et al. 2021; Bechavod et al. 2021]). The most frequently used norms are $p = 1$ and $p = 2$.

(ii) (separable) $\mathsf{cost}(x, \hat{x}) := c(\hat{x}) - c(x)$ (e.g., [Hardt et al. 2016; Hu et al. 2019]). These cost functions are suitable for settings where achieving each feature has a certain cost, but this is independent of which feature the agent started from.

The vast majority of the literature assumes that (given the aforementioned utilities) the agents are *best responding* to f, i.e., that $\hat{x}(f) = \arg\max_{x' \in \mathcal{X}} u(x, x'; f)$. There are some notable exceptions to this assumptions which we highlight in Section 3. Finally, most of the literature (except a few, e.g., [Dong et al. 2018]) assumes that *all* agents can respond strategically. For example, in classification, even agents with $y = 1$ may want to strategize, if they know that the classification rule f will classify them as 0.

Remark 2.1. In general, we assume that the principal knows the agents' value and cost functions (including δ); they are only missing the original point x and can never fully learn it. To be more specific, given the value and cost functions, the reported $\hat{x}_t$ and the y, the principal *cannot* reverse engineer the original x. There are a couple of works that focus on restricted strategic classification settings where δ is unknown, but the principal can still learn robust decision rules (see Section 3 for details).

Continuous Adaptation vs Manipulation Graph. Some works move away from the continuous[4] model of strategic adaptation. Instead, they introduce the idea of a *manipulation graph* (e.g., [Ahmadi et al. 2023]). In incentive-aware learning with manipulation graphs, the assumption is that there exists a graph $G(\mathcal{X}, E)$ to capture all possible manipulations. In graph G each node corresponds to a different feature vector and each edge $e = (x, x') \in E$ captures the manipulation from x to x'. The cost function then $\mathsf{cost}(x, x')$ is defined as the sum of costs to move from x to x', if such a path exists in G. We will highlight which works use manipulation graphs instead of continuous adaptation in the coming sections.

Full vs Partial Information about the Principal's Algorithm. We have so far assumed that the agent has *full* knowledge of f (or f_t) at the time of choosing their best response.[5] Although this is a useful assumption to understand what solutions are possible in the worst case, in reality it is far from the truth; while agents do exhibit strategic adaptation, they seldom have *full* information about the decision-making rules used. There has been an emerging interest in modeling partial information from the agent side (e.g., [Braverman and Garg 2020; Ghalme et al. 2021; Bechavod et al. 2022; Cohen et al. 2024]), but no single model has prevailed as the canonical one. We highlight these models in the coming sections.

Heterogeneous Agents

Finally, we have so far assumed that there is a single $\mathcal{D}$ representing the entire population and that every agent shares the same utility function. In other words, we have assumed that agents are *homogeneous*. However, this assumption is often unrealistic; for instance, in the context of school admissions, it is unlikely that everyone in the population has the same natural ability to succeed in school or the same capacity to take steps to improve their chances of being admitted.

Agent heterogeneity has been studied primarily in two different forms. First, agents may come from heterogeneous populations (i.e., their features and labels may originate from different distributions e.g., [Milli et al. 2019; Hu et al. 2019]). Second, agents may have different abilities to adapt to the decision rule that the principal is using (either because of different cost functions e.g., [Milli et al. 2019; Hu et al. 2019]) or because of different understanding of the decision rule (in the case of partial information) (e.g., [Bechavod et al. 2022]). We discuss heterogeneous agents in Section 5.

[4]The literature sometimes refers to this type of strategizing as "ball manipulation".

[5]Historically, this is a byproduct of the fact that the original papers modeled the paper as a Stackelberg game. In the Stackelberg games literature, the standard assumption is that the principal announces their strategy at the beginning of the interaction with the agent. This announcement gives them "commitment power" (as it is referred to in that literature).

Remark 2.2. As should be clear by now, this article focuses exclusively on strategic adaptation that occurs in the *feature* space, rather than the *label* or *response variable* space. There have also been a series of works on ML algorithms when the agents can strategically adapt their label (e.g., [Dekel et al. 2010; Chen et al. 2018]) but they are beyond the scope of this article. The aforementioned articles take a "robustness" perspective.

One final note: the terminology introduced in this section will be used throughout the following sections to describe each paper. This consistent terminology is intended to help the reader develop a clear mental framework for understanding the types of results obtained for each model variant of incentive-aware learning.

3. ROBUSTNESS PERSPECTIVE MAIN RESULTS

We begin our exposition with the *robustness* perspective. In this framework, the principal seeks to learn the most accurate decision-making rule (as defined in Equation (1)) that maps agent features to a score or classification label, thereby minimizing their loss. Simultaneously, agents strategically manipulate the data they submit to the decision-making rule in an effort to "game" the system. We first examine the offline/batch and online learning settings in Sections 3.1 and 3.2, respectively, focusing on scenarios where agents have full knowledge of the principal's decision-making rule. Subsequently, in Section 3.3, we explore settings where agents have only *partial* information about the principal's decision-making rule. Finally, we conclude this section by discussing cases where agents are *not* individually rational when selecting their misreports, $\hat{x}(f)$, in Section 3.4.

3.1 Offline and Batch Learning Setting

[Hardt et al. 2016] introduced the problem of "strategic classification" in the *offline* setting and formulated it as a Stackelberg game. In their framework, the population of agents is assumed to be *homogeneous*, with each agent aiming to maximize their probability of being classified as $+1$ while incurring a cost for doing so. The principal, on the other hand, wants to design a classifier that converges to the offline optimal in terms of "accuracy" (as defined in Equation (1)) for the $0-1$ loss. The agents are assumed to have *full* information about the classification rule and are best-responding to it. The authors show that for agents with separable cost functions, it is possible to design efficient and nearly optimal classifiers, even for concept classes that are computationally hard to learn. Their theoretical framework further includes impossibility results for learnability when the agents have *general* cost functions, illustrating the fundamental challenges of achieving classification robustness against strategic behavior.

Working in the *offline* or *batch* setting with a *homogeneous* population of agents, [Levanon and Rosenfeld 2021] introduce the notion of *strategic empirical risk minimization* (strategic ERM) as an approach for designing strategy-robust decision rules for the principal. At a high level, the authors propose a "smoothed" version of the strategic classification problem, incorporating the agents' best-response behavior as a function of the decision rule f into the optimization process for f. While the paper does not provide theoretical guarantees, it includes a series of experiments demonstrating how strategic ERM might perform in practice. However,

the assumption of a "smoothed" version of the problem has limitations from a real-world modeling perspective. As noted by several works (e.g., [Dong et al. 2018; Chen et al. 2020]), the motivating settings for strategic classification often make it infeasible to identify a "smooth" loss function for the principal once the agents' best-response behavior is incorporated.

Still working within the ERM paradigm, in the *offline* learning setting and drawing intuition from traditional PAC learning [Valiant 1984], there has also been interest in a PAC version of incentive-aware learning, i.e., given a set of points that have been strategically modified, identify the complexity of finding a classification function that is $\varepsilon(\eta)$-optimal (according to Equation (1)) with high probability at least $1-\eta$. This version of the problem was introduced by [Zhang and Conitzer 2021]. The authors assume that the agents can best-respond according to a *reporting structure* which maps original features to manipulated ones.[6] Moreover, they assume that the principal is facing a *homogeneous* population of agents. The paper first shows that the vanilla ERM (i.e., the one ignoring incentives) has poor performance in strategic settings.[7] Subsequently, they show that a version of *strategic empirical loss* can obtain nearly optimal sample complexity bounds. To construct their strategic empirical loss, the authors take a "worst-case perspective"; for each reported point, they substitute it with the worst-possible original point it could have originated from.[8]

In a similar vein, [Lechner and Urner 2022] study the learnability of general concept classes with a new class of loss functions called *strategic loss* (which is used as a proxy hypothesis class for the principal). In their setting, the agents can manipulate according to a manipulation graph. The strategic loss is a discrete loss function which takes a value of 1 every time that either $f(x) \neq y$ (i.e., incorrect classification) or $f(x) = 0$ but there exists a point x' such that x' is a reachable misreport from x and $f(x') = 1$, and 0 otherwise. This new loss function aims to not only account for accuracy but also, for the societal burden that is induced when the agents fool the classifier.

[Sundaram et al. 2023] take incentive-aware PAC learnability one step further; the agent population is now *heterogeneous* (i.e., the cost function is the same across agents, but each agent may have a different γ in their value function), the principal does *not* know the agents' value functions, but the principal has access to a training dataset that is un-manipulated (i.e., the principal can see some original x's). The key contribution of the work is the introduction of the *Strategic VC-Dimension* (the strategic analogue of VC-dimension), which quantifies learnability in settings where test data is strategically manipulated based on *heterogeneous* agents. The authors subsequently characterize the statistical and computational limits of strategic linear classification. This study also explores the role of *randomization* in improving accuracy under strategic manipulation. We expand on the role of *randomness* in

[6] This can be considered as part of the general "manipulation graph"-type of cost functions.

[7] For the online setting, a slightly stronger result of two-way incompatibility between regular and strategic settings was obtained by [Chen et al. 2020]. Specifically, the authors show that there exist classification settings for which every no-external regret algorithm incurs linear Stackelberg regret and vice versa.

[8] A version of this technique was also used in [Chen et al. 2020], albeit for the online version of the problem.

strategic classification settings in Section 3.3.

[Rosenfeld and Rosenfeld 2024] focus on learning a *linear* classifier (the principal has access to a set of un-manipulated data at training time), the agents have a $0-1$ value function, and L_p-norm cost function. Importantly, the authors assume that δ (i.e., the cost function) is *not* known by the principal but is the same[9] across all agents; yet, the principal still needs to learn a classification rule that converges to the optimal one. The authors take a robust optimization approach, by minimizing the worst-case risk over a family of costs which includes the target (unknown) cost. They do so, because as they show, if the principal has to commit to a single fixed cost for their risk minimization problem, then ERM can never provide a non-trivial data-independent guarantee (unless the assumed single fixed cost was precisely correct). As for the ERM, the authors consider a type of *hinge* loss, that is appropriately expanded in order to include the uncertainty induced by the unknown cost function. The main result of the paper is an efficient iterative algorithm that converges to the minimax optimal solution with rate $\tilde{O}(1/\sqrt{T})$, where $\tilde{O}(\cdot)$ hides polylogarithmic terms, and T is the number of the algorithm's iterations.

3.2 Online Learning Setting

The online learning version of strategic classification was first studied by [Dong et al. 2018]. In their paper, the authors provide linear strategic classification algorithms with sublinear Stackelberg regret (see Equation (2) against a *homogeneous* population of agents with *linear* values (i.e., the agents care about maximizing their distance from the classifier, while being labeled as $+1$ by it). To give an overview of their approach, let w_t be the normal vector corresponding to classifier f_t for each round $t \in [T]$, i.e., $f_t(x) := w_t^\top x$. The main result of the paper is to find the sufficient conditions on the agents' $\hat{x}(w_t)$ such that $\ell(w_t, \hat{x}_t(w_t))$ is *convex* in w_t, when $\ell(w_t, \hat{x}_t(w_t))$ is either the hinge or logistic loss. This task boils down to identifying the sufficient conditions on the agents' cost function in order for $\ell(w_t, \hat{x}(w_t))$ to be convex in w_t. Convexity is desired, since if $\ell(w_t, \hat{x}_t(w_t))$ is convex in w_t, then the principal can apply any off-the-shelf bandit convex optimization algorithm and obtain sublinear Stackelberg regret. The paper obtains improved regret bounds under the assumptions that all agents with $y_t = 1$ are non-strategic.

But what happens when $\ell(w_t, \hat{x}_t(w_t)$ is not a convex function of w_t? In an effort to answer this question in a general way, [Chen et al. 2020] studied online learning of linear classifiers in the following setting: the agents have a discrete value for passing the classifier (i.e., they obtain a value of 1 for passing the classifier and 0 otherwise), their cost function is δ-bounded (i.e., $\|\hat{x}_t(f_t) - x_t\| \leq \delta, \forall t \in [T]$), and the learner cares about the $0-1$ loss. Importantly, the results of the paper do not require the agents to *rationally* best-respond; instead, knowing that the $\hat{x}_t(f_t)$ satisfy the constraint that $\|\hat{x}_t(f_t) - x_t\|_2 \leq \delta$ is enough. The paper provides a nearly tight algorithm that dynamically and adaptively partitions the space of feasible classifiers for the principal as new agents arrive. The final Stackelberg regret bound depends on the *instance* of datapoints $\{(x_t, y_t)\}_{t\in[T]}$ that nature chooses. The key tricks that the authors use is that when the principal sees a reported point $\hat{x}_t(f_t)$, then they know for sure that the true x_t lies inside a ball B, where

[9]This is the main difference with the *model* of [Sundaram et al. 2023].

$B := \{x \in \mathcal{X} : \|x - \hat{x}_t(f_t)\|_2 \leq \delta\}$. The final trick is to observe that given this information and the fact that the learner cares about the $0-1$ loss, then the principal can obtain perfect information about the loss that would have been incurred in that round t if the same agent at round t were to best respond to some other normal vectors w for which $\|\hat{x}_t(w) - \hat{x}_t(f_t)\|_2 \leq 2\delta$. The theoretical analysis of the algorithm requires knowing the magnitude of the agents' manipulation (δ) and access to a carefully crafted oracle that can provide some extra information to the principal about the structure of the agents' unmanipulated data.

The aforementioned paper trades efficiency for generality. When the sequence of data $\{(x_t, y_t)\}_{t\in[T]}$ chosen by nature is *separable* by a margin, [Ahmadi et al. 2021] introduce a variant of the Perceptron algorithm, called the *Strategic Perceptron*, which is *computationally efficient* and converges to a maximum-margin classifier while making a bounded number of mistakes. The upper bound on the number of mistakes depends on the margin of the original, unmanipulated data and the agents' strategizing power. The Strategic Perceptron is analyzed under the assumption that agents incur either L_1 or L_2 costs when misreporting from x to $\hat{x}$, and are rationally best-responding. Notably, the paper shows how to leverage the structure of the agents' utility function together with the fact that the agents are rationally best responding to establish bounded mistake guarantees *even when* the magnitude of the manipulation cost is *not* known to the principal a priori — a result that was not achievable in [Chen et al. 2020].

Next, we transition from models of continuous strategic adaptation to models where agents determine their $\hat{x}_t$ based on a manipulation graph, highlighting the work of [Ahmadi et al. 2023]. This setting generalizes the frameworks of [Zhang and Conitzer 2021] and [Lechner and Urner 2022] to the online setting. The paper demonstrates that, unlike in the non-strategic classification setting, the vanilla Halving algorithm may incur an infinite number of mistakes. To address this, the authors propose a general algorithm for the strategic setting with a mistake bound of $O(\Delta \ln(|\mathcal{H}|))$, where Δ is the degree of the manipulation graph and $\mathcal{H}$ is the (known) class of the target function. Furthermore, the paper extends the algorithm to the agnostic learning setting.

Adopting a similar perspective of testing the limits of strategic learnability, [Cohen et al. 2024] and [Ahmadi et al. 2021] investigate whether the learnability of a concept class implies its strategic learnability. They essentially show that every learnable function class remains learnable even when data is strategically manipulated. Both works model the agents' feasible manipulations using manipulation graphs and consider scenarios where the graph is either fully known or only partially known to the principal. [Ahmadi et al. 2021] introduce the "strategic Littlestone dimension," which captures the complexity of the agents' manipulation graph and the hypothesis class. Both papers analyze strategic learnability across multiple variations of the baseline strategic classification model. Finally, [Shao et al. 2024] study learnability in terms of mistake bounds and sample complexity when agents' manipulations are *heterogeneous*. They consider both continuous adaptations and manipulation graphs. As for the principal, they assume that some knowledge of x_t is available either before choosing the classification rule f_t or immediately afterward.

In a slightly different setup, [Harris et al. 2023] consider an online setting where

at each round the principal commits to a function $f_t : \mathcal{X} \to \{0, 1\}$, the agents can strategically adapt within a ball of radius δ of their true datapoint x_t, and the reward that the principal receives is linear in the agent's unmodified context; more concretely, for each decision $\alpha \in \{0, 1\}$ the reward of the principal for a context x_t is: $r_t(\alpha) = \theta_\alpha^\top x_t + \varepsilon$, where θ_α is a $d-$dimensional vector. The authors assume that the principal has "apple tasting" feedback, i.e., the principal can observe $r_t(\alpha)$ only when $\alpha = 1$ (which in turn, is decided by the function f_t). The authors present algorithms that actually incentivize agents to be *truthful* (i.e., report x_t without any manipulation) while achieving sublinear regret.

3.3 Partial Information about the Principal's Algorithm

So far, we have primarily assumed that the principal commits to a *deterministic* rule and that agents fully observe this rule. [Braverman and Garg 2020] were the first to highlight the role of *randomness* in the principal's classifier and the impact of *noise* in the agents' features on the outcomes of the strategic classification game. The paper demonstrates that to maximize accuracy (as defined by Equation (1)), the principal may *need* to employ randomized rules. This result creates an intriguing policy dilemma: on the one hand, randomized rules may be necessary to achieve optimal accuracy; on the other hand, their deployment can be legally problematic. Interestingly, the paper shows that introducing (or having inherently) noisier signals for the agents' features can improve both accuracy and fairness in equilibrium across different subpopulations.

[Ahmadi et al. 2023] also explore the role of randomness in strategic classification, focusing on its impact on learnability. They consider two sources of randomness. In the first, the principal commits to a probability distribution over classifiers, thereby inducing certain probabilities of classification as $+1$ for agents. In the second, the principal commits to a probability distribution over classifiers, nature (which may adversarially select the next x_t) responds to this distribution, and the chosen agent x_t best responds to the *realized* classifier. The second model is more *transparent* to the agents than the first and enables the principal to design algorithms with improved regret guarantees.

If the principal has the choice between a transparent and an "opaque" classifier, which approach minimizes prediction error? [Ghalme et al. 2021] address this question in the setting of [Hardt et al. 2016] (i.e., offline, homogeneous population of agents, etc.). They define the *price of opacity* as the difference in prediction error when agents respond to a fully transparent classifier f versus an opaque rule $\hat{f}$. The paper studies the conditions under which the price of opacity can be positive or negative. Consistent with the theory of Stackelberg games, revealing f (or allowing it to be fully anticipated or deduced from $\hat{f}$) can sometimes benefit the principal, as it enables them to precisely predict how agents will react.

[Cohen et al. 2024] introduce a Bayesian classification setting, where the principal gradually reveals information about the classification rule. In this model, agents share a common distributional prior over the classifier used by the principal and best respond by maximizing their expected utility. The principal, in turn, can strategically release partial information about the classifier over time. The authors show how to release this information carefully to ensure that truly qualified agents

(i.e., $y_t = +1$) can pass the classifier while preventing unqualified agents from gaining sufficient information to successfully strategize and game the system.

Finally, [Bechavod et al. 2022] study a setting where agents acquire information about the classifier through "peer learning." The primary focus of this work is on the fairness implications of information discrepancies across different subpopulations. Therefore, we defer a detailed discussion of this work to Section 5.

3.4 Beyond Rational Best-Response Agents

So far, we have focused on settings where agents best-respond to the principal's rule. We now shift our attention to scenarios where agents do *not* precisely best-respond.

Although the results in [Chen et al. 2020] hold for *any* agent manipulation within δ of the true data point, [Jagadeesan et al. 2021] formalize alternative models for agent behavior that deviate from exact, rational best response. The authors demonstrate the brittleness of standard strategic classification algorithms when agents do not strictly adhere to the assumed best-response model. To address this, they identify a set of desiderata for agent responses that ensure algorithm stability and propose the *noisy response* model. In this model, agents best respond to a noise-perturbed version of the decision rule, inspired by the principles of smoothed analysis [Spielman and Teng 2009].

[Ebrahimi et al. 2024] study the role of behavioral biases in agents' responses within strategic classification settings. Specifically, they consider agents who, when evaluating the value of passing the classifier, *weigh* the classifier's features according to their own biases. The paper analyzes a homogeneous population of agents who can incur a cost of up to B for misreporting. It identifies cases where agents overshoot or undershoot the classifier's boundary due to their biased perceptions of the classifier's feature weights.

[Lechner et al. 2023] examine settings where the principal faces two sources of uncertainty regarding the agents' responses. First, agents are not required to rationally best-respond and are instead permitted to use any *feasible* response that enables them to fool the classifier. Second, the principal does not have full knowledge of the agents' manipulation graph but only knows the general family to which it belongs. Focusing on strategic loss, the authors study the learnability of both proper and improper learning under these assumptions. Their key result is that it is possible to learn an almost-optimal classifier in terms of strategic loss, even without precise knowledge of the manipulation graph.

[Cohen et al. 2024] explore the effects of partial knowledge of the manipulation graph on learnability. They show that when the principal knows only the general family of graphs to which the manipulation graph belongs, they can achieve nearly tight bounds on both sample complexity and regret. Furthermore, the difference in learning complexity between the fully-known and partially-known graph settings is (roughly) logarithmic in the size of the graph family.

Finally, [Ahmadi et al. 2024] also assume that the principal knows only the *family* of graphs to which the agents' manipulation graph belongs. They derive a regret bound that is approximately optimal for certain instances. Additionally, they extend their results to a setting where each agent may have a different manipulation graph, provided all graphs belong to the same family. This generalized setting is referred to as the "agnostic" case.

4. IMPROVEMENT & CAUSALITY PERSPECTIVE MAIN RESULTS

Oftentimes strategic adaptation to algorithmic decision-making rules may lead to genuine improvement for the individuals; for instance, paying-off prior debt as a means of increasing your credit score actually helps you become more creditworthy. To state this in the language of incentive-aware learning, this means that the agents' label y can change when they switch from their true x to the strategically manipulated $\hat{x}$. This section focuses on settings where strategic adaptation can lead to both gaming and actual improvement.

4.1 Improvement & Recourse

According to the "improvement"/"recourse"[10] perspective, any strategic adaptation results in genuine improvement for individuals; that is, when a data point changes from x to $\hat{x}$, it holds that $h^\star(x) < h^\star(\hat{x})$.

[Kleinberg and Raghavan 2020] introduce one such model where agent "manipulations" result in changes to the underlying features, which can constitute genuine improvement for the agents. Their primary result shows that simple linear mechanisms suffice to incentivize genuine improvement in settings where the principal interacts with a single agent. [Harris et al. 2021] extend the model of [Kleinberg and Raghavan 2020] to settings where agents achieve improvements over a sequence of rounds, i.e., agents transition through different states over time as they respond to the principal's rule.

[Alon et al. 2020] generalize the single-agent setting of [Kleinberg and Raghavan 2020] to a multi-agent framework. In their model, all agents share the same initial feature representation, but their ability to manipulate (quantified by their manipulation costs) differs.

[Haghtalab et al. 2020] also study multi-agent settings, focusing on designing evaluation mechanisms that maximize population-wide quality scores when agents can strategically alter their features at a cost. Their model differs from [Alon et al. 2020] in that agents can have different initial feature representations. The authors analyze two specific classes of mechanisms: linear mechanisms and linear threshold mechanisms. For linear mechanisms, they show that the optimal strategy corresponds to projecting the true quality function onto the observable feature space, which is computationally efficient. For linear threshold mechanisms, they develop approximation algorithms, including a constant-factor approximation algorithm under smooth feature distributions, that balance the trade-offs between incentivizing improvements and maximizing welfare. The paper further considers scenarios where the feature distribution is unknown and provides sample-complexity guarantees for learning optimal mechanisms.

[Tsirtsis and Gomez Rodriguez 2020] explore the design of optimal decision-making policies and counterfactual explanations in incentive-aware learning. They

[10]The term "recourse" comes from the traditional ML literature. Loosely speaking, algorithmic recourse refers to the ability of individuals to reverse negative decisions made by algorithms through counterfactual explanations provided alongside the decision. A substantial body of work exists on algorithmic recourse (see, e.g., [Karimi et al. 2020] for a survey), but it is beyond the scope of this article. Here, we focus specifically on the effects of strategic adaptation on algorithmic recourse.

model this problem as a Stackelberg game, where decision-makers provide *counterfactual explanations*—guidelines on how agents can change their features—and agents respond strategically to maximize their benefit. Unlike the standard Stackelberg game for incentive-aware learning, where the decision rule is announced, here the principal announces counterfactual explanations. The authors show that optimizing the set of counterfactual explanations for a fixed decision policy is NP-hard but can be addressed using approximation algorithms that leverage the problem's submodularity. They further extend the problem to jointly optimize both the decision policy and explanations, reducing it to a non-monotone submodular maximization problem solvable with approximation guarantees. Additionally, the paper incorporates diversity constraints to ensure equitable distribution of explanations across populations.

Finally, [Bechavod et al. 2022] study the "improvement" perspective when the principal's decision rule is not fully known to the agents. Their work focuses on the effects of information discrepancies across different subpopulations and is therefore discussed in Section 5.

4.2 Causality

How can we distinguish between agent actions that lead to genuine improvement versus those that constitute mere gaming? As [Miller et al. 2020] observe, designing "good" incentive-aware decision-making rules—rules that incentivize actions leading to genuine improvement while disincentivizing gaming—is equivalent to identifying the causal model underlying the setting (i.e., performing causal inference). Their work was the first to formalize this connection, introducing causal graphs to study how certain agent features affect (or do not affect) the target variable y.

Building on the theme of causality in incentive-aware learning, [Shavit et al. 2020] study incentive-aware linear regression, where the decision-maker seeks to optimize one of three objectives: (1) Agent Outcome Maximization (incentivizing agents to improve their outcomes), (2) Prediction Risk Minimization (ensuring accurate prediction of post-gaming outcomes), and (3) Parameter Estimation (accurately estimating the causal parameters of the outcome-generating process). The authors propose efficient algorithms for each objective in a realizable linear setting, leveraging the ability to test and observe agent responses to decision rules—effectively performing causal interventions. This ability to perform interventions makes their setting more tractable compared to [Miller et al. 2020]. Additionally, they address challenges such as omitted variable bias and interactions between observed and hidden features, which can undermine naive regression approaches. Extending beyond linear regression, [Horowitz and Rosenfeld 2023] study (agnostic) incentive-aware classification under causality with the goal of improving the principal's accuracy.

In concurrent and independent work, [Bechavod et al. 2021] explore incentive-aware linear regression and demonstrate how agents' strategic behavior can facilitate the learning of causal variables. The authors propose a batch-based retraining approach that iteratively updates the regression model, leveraging agents' strategic modifications to improve predictive accuracy while incentivizing genuine improvement in features. They prove that this dynamic interaction enables the principal to accurately recover the true regression parameters over time, even when features are correlated. As a result, the principal can both incentivize genuine improvement

and improve the robustness of the decision model.

Finally, [Ahmadi et al. 2022] study incentive-aware classification under a causal model, addressing both discrete strategic adaptation (via manipulation graphs) and continuous adaptation. For the general discrete model, the authors design efficient algorithms to maximize true positives while ensuring no false positives, thus guaranteeing that only genuinely qualified agents are classified positively. They further show that the problem of selecting criteria to maximize true positives while allowing even a bounded number of false positives becomes NP-hard. In the continuous adaptation (linear) model, they develop algorithms to determine whether a linear classifier exists that classifies all agents accurately while incentivizing all improvable agents to become qualified.

4.3 Performativity

Before we conclude the section on improvement and causality in incentive-aware ML settings, we briefly touch on the literature on *performative prediction* [Perdomo et al. 2020]. Performative prediction is another framework to explain and reason about how predictions, when used to inform decisions, influence the outcomes they aim to predict. The authors develop a risk minimization framework and propose a new equilibrium notion called performative stability. Roughly speaking, this notion ensures that predictions are calibrated not to past data but to the outcomes they induce. The paper presents necessary and sufficient conditions for retraining algorithms to converge to performatively stable solutions with near-minimal loss. The main distinction between performative prediction and the other models that we highlight in this survey is that performative prediction uses certain smoothness assumptions on the way that original points x leads to strategically adapted points $\hat{x}$, instead of focusing on the agents' utility functions.

5. FAIRNESS MAIN RESULTS

Most (if not all) of the papers discussed so far in this article have focused on a homogeneous population of agents with which the principal is interacting. However, when the principal is interacting with a *heterogeneous* population of agents, with (potentially) different abilities to strategize and different qualifications, then optimizing for the desiderata of robustness-to-gaming or accuracy may have disparate downstream effects to the different subpopulations.

[Hu et al. 2019] and [Milli et al. 2019] independently and concurrently initiated the study of the disparate downstream effects of designing robust-to-gaming classifiers to different subpopulations. [Milli et al. 2019] defined the *social burden* of a classifier as the aggregate of the minimum cost an individual needs in order to be classified as a $+1$. For example, for agents with $y_t = +1$, a high social burden means that it is very costly for the agents to obtain their correct classification. The authors prove a general trade-off between principal's accuracy and agent utility. They also prove that when agents incur cost as a consequence of a principal making their classifier robust to strategic behavior, the costs can disproportionally fall on the disadvantaged subpopulations.

In a similar theme, [Hu et al. 2019] study negative externalities of strategic classification, and show that the Stackelberg equilibrium classifier leads to only false negative errors on the disadvantaged subpopulation but only false positives on the

advantaged population. Not only that, but they also show that providing a cost subsidy (whose goal is to counterbalance this the difference in false negatives and false positives from each subpopulation) can *actually* lead to worse outcomes for *everyone* in the game.

Focusing on the goal of group fairness, [Estornell et al. 2023] explores the unintended consequences of using fairness-aware algorithms in environments where agents can strategically manipulate their features to achieve better outcomes. While fairness in algorithmic decision-making is typically aimed at ensuring equitable treatment across demographic groups, the paper identifies a phenomenon called "fairness reversal". This occurs when a fairness-driven classifier (designed to equalize outcomes between groups) becomes less fair than a conventional accuracy-focused classifier due to strategic feature manipulation by agents. The authors empirically demonstrate this phenomenon using benchmark datasets and attribute it to the selectivity of fair classifiers, which achieve fairness by excluding individuals from the advantaged group rather than including more from the disadvantaged group. They prove that increased selectivity is a sufficient, and in some cases necessary, condition for fairness reversal. They further show that fairness reversal does not occur when fairness is achieved through inclusiveness, where the classifier broadens access to the disadvantaged group.

The focus of the aforementioned works was on fairness in terms of classification accuracy. Lately, some works have started considering fairness in terms of improvement or recourse ability. [Gupta et al. 2019] address fairness in terms of *recourse*, i.e., the effort required to reverse a negative classification, across demographic groups. Mathematically, recourse is measured as the distance from an individual's features to the decision boundary of a classifier. The paper introduces a new approach to regularize classifiers, minimizing disparities in recourse while maintaining predictive accuracy. It extends prior work on linear classifiers to more complex settings, including non-linear models and model-agnostic scenarios, where the decision boundary is not explicitly known. For the model-agnostic setting, the paper assumes that the agents have black-box access to the classifier, rather than the full mathematical formulation.

[Bechavod et al. 2022] study how disparities in information about decision rules affect the ability of agents from different sub-populations to improve their outcomes in strategic learning contexts. Unlike most traditional models that assume agents fully know decision rules, this work focuses on scenarios where decision rules are not fully known originally, and agents infer them based on their peers' experiences, creating group-specific information levels; they refer to this process as "peer learning". The study reveals that even when decision rules are optimized to maximize welfare, disparities in information and effort costs can lead to some sub-populations experiencing a decline in quality ("negative externality"). However, under specific conditions (e.g., proportional costs across groups or minimal information overlap — measured through an "information overlap proxy" metric — across groups) negative impacts can be mitigated.

[Ahmadi et al. 2023] study the problem of designing short-term goal structures to incentivize agents with varying abilities to improve their skills or capacities-for-improvement. It proposes two models: (1) the common improvement capacity

model, where all agents share the same improvement limit, and (2) the individualized improvement capacity model, where agents have personalized improvement limits. The authors develop algorithms to optimize the placement of target skill levels (i.e., goals) to maximize social welfare (i.e., total improvement across all agents) and ensure fairness among groups. One challenge they address is the non-monotonic nature of social welfare, where adding new target levels may unintentionally reduce overall improvement. Finally, they present an extension for the case where the principal has sample access to the available data when designing the classifier.

6. CONCLUSION

The purpose of this article has been to provide a gentle introduction to the exciting area of incentive-aware ML. We categorized the existing research into *robustness, fairness*, and *improvement/causality perspectives*, and we highlighted the diverse approaches and objectives within each domain. We outlined some of the foundational models and theoretical frameworks for understanding strategic adaptation, from offline and online learning settings to causal perspectives, and we emphasized the complexities introduced by agent heterogeneity and partial information.

There have also been a handful of topics related to incentive-aware ML settings that we did not touch upon, as they did not directly fit under one of our three outlined perspectives. Examples include: [Zrnic et al. 2021] who study how the Stackelberg game (and its outcomes) change when the principal and the agent (termed "leader" and "follower" in their paper) alternate in order; papers on econometrics for strategic agents (e.g., [Harris et al. 2022; Harris et al. 2024]); and papers focusing on agents that can choose to not participate in the algorithmic decision making process, if that is aligned with their utility maximization (e.g., [Krishnaswamy et al. 2021], [Horowitz et al. 2024]).

For all the excitement surrounding this research area, there is one question that seems as pressing as ever.

What comes next for the literature on incentive-aware ML?

In the author's view, there are two primary paths for the future of incentive-aware ML. The first path is the more well-established and widely explored. There remain myriad settings requiring theoretical analysis of the interactions between individuals and a decision-making principal. For example, how do information discrepancies about the principal's algorithm across different subpopulations affect their abilities to genuinely improve their outcomes versus merely game the system? Are there properties of "interpretable" decision-making algorithms that can provably incentivize genuine improvement rather than gaming? Developing new models and providing provable guarantees for these questions will help solidify the theoretical foundations of incentive-aware ML.

The second path is less charted and relatively unexplored, particularly from a theorist's perspective. Although examples of individuals strategizing and adapting to algorithmic decision-making rules are abundant, incentive-aware ML still needs to identify a *concrete* application domain where the insights gained from theoretical advancements can *actually be applied*. Such a domain would allow incentive-aware algorithms to be deployed and evaluated against other "robust" algorithms.

This approach differs from the path the literature has predominantly taken. To illustrate this distinction, consider the steps required to apply theoretical insights from incentive-aware ML to a practical domain, such as recommendation systems (RecSys).[11] To apply these insights effectively in the RecSys domain, we would need to address several questions: Do users "strategize" with their data (see e.g., [Haupt et al. 2023])? What utility function are they optimizing for? What does it mean for users to have "partial" information about the RecSys? What specific interventions can the RecSys implement to mitigate inequalities between different user subpopulations?

Identifying such a concrete application domain would enable the foundational results in this field to be translated into actionable insights, driving meaningful, real-world change. The author is optimistic about the potential of the next generation of incentive-aware ML research to bridge this gap and create significant societal impact.

REFERENCES

Ahmadi, S., Beyhaghi, H., Blum, A., and Naggita, K. 2021. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 6–25.

Ahmadi, S., Beyhaghi, H., Blum, A., and Naggita, K. 2022. On classification of strategic agents who can both game and improve. In *3rd Symposium on Foundations of Responsible Computing, FORC 2022, June 6-8, 2022, Cambridge, MA, USA*, L. E. Celis, Ed. LIPIcs, vol. 218. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 3:1–3:22.

Ahmadi, S., Beyhaghi, H., Blum, A., and Naggita, K. 2023. Setting fair incentives to maximize improvement. In *4th Symposium on Foundations of Responsible Computing, FORC 2023, June 7-9, 2023, Stanford University, California, USA*, K. Talwar, Ed. LIPIcs, vol. 256. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 5:1–5:22.

Ahmadi, S., Blum, A., and Yang, K. 2023. Fundamental bounds on online strategic classification. In *Proceedings of the 24th ACM Conference on Economics and Computation*. 22–58.

Ahmadi, S., Yang, K., and Zhang, H. 2024. Strategic littlestone dimension: Improved bounds on online strategic classification. *arXiv preprint arXiv:2407.11619*.

Alon, T., Dobson, M., Procaccia, A., Talgam-Cohen, I., and Tucker-Foltz, J. 2020. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 1774–1781.

Bechavod, Y., Ligett, K., Wu, S., and Ziani, J. 2021. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1234–1242.

Bechavod, Y., Podimata, C., Wu, S., and Ziani, J. 2022. Information discrepancy in strategic learning. In *International Conference on Machine Learning*. PMLR, 1691–1715.

Björkegren, D., Blumenstock, J. E., and Knight, S. 2020. Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

Braverman, M. and Garg, S. 2020. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference)*, A. Roth, Ed. LIPIcs, vol. 156. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 9:1–9:20.

Chang, T., Warrenburg, L., Park, S.-H., Parikh, R. B., Makar, M., and Wiens, J. 2024. Who's gaming the system? a causally-motivated approach for detecting strategic adaptation. *arXiv preprint arXiv:2412.02000*.

[11] Another promising application domain is the health insurance industry, as recently discussed in [Chang et al. 2024].

Chen, Y., Liu, Y., and Podimata, C. 2020. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems 33*, 15265–15276.

Chen, Y., Podimata, C., Procaccia, A. D., and Shah, N. 2018. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 9–26.

Cohen, L., Mansour, Y., Moran, S., and Shao, H. 2024. Learnability gaps of strategic classification. *arXiv preprint arXiv:2402.19303*.

Cohen, L., Sharifi-Malvajerdi, S., Stangl, K., Vakilian, A., and Ziani, J. 2024. Bayesian strategic classification. *arXiv preprint arXiv:2402.08758*.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. 2019. The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics 11,* 3, 382–423.

Dekel, O., Fischer, F., and Procaccia, A. D. 2010. Incentive compatible regression learning. *Journal of Computer and System Sciences 76,* 8, 759–777.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 55–70.

Dranove, D., Kessler, D., McClellan, M., and Satterthwaite, M. 2003. Is more information better? the effects of "report cards" on health care providers. *Journal of political Economy 111,* 3, 555–588.

Ebrahimi, R., Vaccaro, K., and Naghizadeh, P. 2024. The double-edged sword of behavioral responses in strategic classification: Theory and user studies. *arXiv preprint arXiv:2410.18066*.

Estornell, A., Das, S., Liu, Y., and Vorobeychik, Y. 2023. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 389–399.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., and Rosenfeld, N. 2021. Strategic classification in the dark. In *International Conference on Machine Learning*. PMLR, 3672–3681.

Gonzalez-Lira, A. and Mobarak, A. M. 2019. Slippery fish: Enforcing regulation under subversive adaptation. Tech. rep., IZA Discussion Papers.

Greenstone, M., He, G., Jia, R., and Liu, T. 2022. Can technology solve the principal-agent problem? evidence from china's war on air pollution. *American Economic Review: Insights 4,* 1, 54–70.

Gupta, V., Nokhiz, P., Roy, C. D., and Venkatasubramanian, S. 2019. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Z. 2020. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, Ed. ijcai.org, 160–166.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 111–122.

Harris, K., Agarwal, A., Podimata, C., and Wu, Z. S. 2024. Strategyproof decision-making in panel data settings and beyond. *ACM SIGMETRICS Performance Evaluation Review 52,* 1, 69–70.

Harris, K., Heidari, H., and Wu, S. Z. 2021. Stateful strategic regression. *Advances in Neural Information Processing Systems 34*, 28728–28741.

Harris, K., Ngo, D. D. T., Stapleton, L., Heidari, H., and Wu, S. 2022. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*. PMLR, 8502–8522.

Harris, K., Podimata, C., and Wu, S. Z. 2023. Strategic apple tasting. *Advances in Neural Information Processing Systems 36*, 79918–79945.

Haupt, A., Hadfield-Menell, D., and Podimata, C. 2023. Recommending to strategic users. *arXiv preprint arXiv:2302.06559*.

Horowitz, G. and Rosenfeld, N. 2023. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*. PMLR, 13233–13253.

Horowitz, G., Sommer, Y., Koren, M., and Rosenfeld, N. 2024. Classification under strategic self-selection. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Hu, L., Immorlica, N., and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 259–268.

Jagadeesan, M., Mendler-Dünner, C., and Hardt, M. 2021. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*. PMLR, 4687–4697.

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Kleinberg, J. and Raghavan, M. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC) 8,* 4, 1–23.

Krishnaswamy, A. K., Li, H., Rein, D., Zhang, H., and Conitzer, V. 2021. Classification with strategically withheld data. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5514–5522.

Lechner, T. and Urner, R. 2022. Learning losses for strategic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7337–7344.

Lechner, T., Urner, R., and Ben-David, S. 2023. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*. PMLR, 18714–18732.

Levanon, S. and Rosenfeld, N. 2021. Strategic classification made practical. In *International Conference on Machine Learning*. PMLR, 6243–6253.

Miller, J., Milli, S., and Hardt, M. 2020. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*. PMLR, 6917–6926.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 230–239.

Pearl, J. 2009. *Causality*. Cambridge university press.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

Rosenfeld, E. and Rosenfeld, N. 2024. One-shot strategic classification under unknown costs. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Shao, H., Blum, A., and Montasser, O. 2024. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems 36*.

Shavit, Y., Edelman, B., and Axelrod, B. 2020. Causal strategic linear regression. In *International Conference on Machine Learning*. PMLR, 8676–8686.

Spielman, D. A. and Teng, S.-H. 2009. Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Communications of the ACM 52,* 10, 76–84.

Sundaram, R., Vullikanti, A., Xu, H., and Yao, F. 2023. Pac-learning for strategic classification. *Journal of Machine Learning Research 24,* 192, 1–38.

Tsirtsis, S. and Gomez Rodriguez, M. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems 33*, 16749–16760.

Valiant, L. G. 1984. A theory of the learnable. *Communications of the ACM 27,* 11, 1134–1142.

Zhang, H. and Conitzer, V. 2021. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5797–5804.

Zrnic, T., Mazumdar, E., Sastry, S., and Jordan, M. 2021. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems 34*, 15257–15269.

Automated Mechanism Design: A Survey

MICHAEL J. CURRY
University of Illinois Chicago
and
ZHOU FAN, YANCHEN JIANG, SAI SRIVATSA RAVINDRANATH, TONGHAN WANG,
DAVID C. PARKES
Harvard University

In this note, we survey automated mechanism design (AMD): the use of computational techniques to solve mechanism design problems. We describe three distinct but overlapping threads of research: an optimization-based paradigm that formulates mechanism design as linear programming, a line of work on sample complexity and learning theory, and the recent trend of *differentiable economics*, which has produced state-of-the-art results on a range of problems by borrowing tools and techniques from modern deep learning.

1. INTRODUCTION

Automated mechanism design (AMD), construed broadly, is the use of computational techniques to find solutions to specific instances of mechanism design problems. The term was introduced in a paper by Conitzer and Sandholm [2003a] just over two decades ago. Since then, AMD has developed into a rich interdisciplinary field integrating methods from across computer science, including optimization, learning theory, and, most recently, deep learning.

There are practical reasons to care about automated mechanism design. Many organizations already use a data-driven approach for pricing and bundling decisions, and even minor improvements can have a large financial impact. For instance, recent work at Yahoo shows a revenue increase of millions of dollars from optimizing a single parameter of their ad auctions [Alcobendas et al. 2024]. Automated mechanism design is a generalization of such optimization: while to our knowledge very general automated mechanism design techniques have not yet been used to field such important mechanisms, the potential seems large.

Researchers can also make use automated mechanism design as a tool for theory. As we will discuss further below, there are many natural and interesting mechanism design problems (most notably revenue-maximizing auction design) that have resisted clean theoretical characterization. Solutions computed using automated mechanism design can act as conjectures and offer insight into problem structure.

In the following sections, as we survey the last two decades of AMD research, we will focus on three main threads. First, we will discuss the earliest approach to AMD, which frames mechanism design as an optimization problem, typically represented as a linear program (LP) (see Section 3). This approach is reasonably general, although limited to finite type spaces, and benefits from decades of advances in LP solvers. However, many mechanism design problems of interest require an unreasonably large LP.

The second thread of research connects automated mechanism design to *learn-*

ing theory (Section 4). Given a class of mechanisms, and information about the distribution of participants in the form of samples, this approach aims to select a mechanism with high estimated performance. Ideally, there is some guarantee of generalization from the set of samples to the true distribution. This is exactly the type of problem learning theory tries to tackle. Conveniently, many interesting classes of mechanisms have structural properties that can be connected back to learning-theoretic properties such as pseudodimension and Rademacher complexity.

The most recent thread of research, known as *differentiable economics* (Section 5), also treats mechanism design as a learning problem—specifically, a *deep learning* problem. In using deep learning as a tool for discovering economic mechanisms, differentiable economics is similar to efforts in the natural sciences to use deep learning for scientific discovery [Wang et al. 2023]. This approach borrows the computational tools of modern deep learning (such as composable, differentiable function approximators and gradient-descent-based optimization) and embraces a deep learning sensibility: it is relatively pragmatic and empirical, willing to relax hard guarantees in order to get better results. Despite somewhat less theoretical grounding, differentiable economics has in practice been remarkably successful, producing state-of-the-art mechanisms for a number of interesting problems.

2. MECHANISM DESIGN BACKGROUND

In this section, we give a brief and fairly standard definition of mechanism design. The purpose of this description is to provide some concreteness and to introduce notation—it is not meant to be an exhaustive or canonical formulation of mechanism design. Then, we discuss a smörgåsbord of basic theoretical results that are used in automated mechanism design.

Appealing to the *revelation principle*, we focus on *direct-revelation mechanisms*, which accept reports from each agent of their preferences over the outcome to be chosen by the mechanism (known as the *type* of the agent). We suppose agent types lie in some space $\mathcal{V}$, with outcomes in space $\mathcal{O}$. We index individual agent types as v_i and, in multi-agent settings, refer to *type profiles* $v = (v_1, \cdots, v_i, \cdots, v_n) = (v_i, v_{-i})$. We will treat types and outcomes as dual, so the welfare enjoyed for a bidder i with type v_i who receives outcome o can be thought of as computing an inner product $v_i \cdot o$. (This is fully general – following Frongillo and Kash [2021], we can consider $\mathcal{O}$ to be the space of distributions over all base outcomes, $\mathcal{V}$ to be functions on this space, and the inner product to be integrating v_i against an o.)

We focus on bidders with *quasilinear utilities*, so that bidder i's utility with type v_i for receiving outcome o and paying a monetary transfer of $p_i \in \mathbb{R}$ is $u(v_i) = v_i \cdot o - p_i$.

Definition 2.1 *Typical Mechanism Design Problem.* A mechanism design problem instance is defined by a distribution P over participant type profiles $v \in \mathcal{V}^n$ for a set of n agents, and a *performance goal* $\mathcal{L}$. The mechanism designer chooses an *allocation rule* $a : \mathcal{V}^n \to \mathcal{O}$ which maps type profiles onto outcomes, and a *payment rule* $p : \mathcal{V}^n \to \mathbb{R}^n$.

When seeking a dominant strategy equilibrium, the mechanism designer must

solve

$$\max_{a,p} \mathbb{E}_{v\sim P}\left[\mathcal{L}(a,p,v)\right] \text{ s.t.}$$
$$a(v_i, v_{-i}) \cdot v_i - p_i(v_i, v_{-i}) \geq a(b_i, v_{-i}) \cdot v_i - p_i(b_i, v_{-i}), \quad \forall v, b_i, i \quad \text{(DSIC)}$$
$$a(v_i, v_{-i}) \cdot v_i - p_i(v_i, v_{-i}) \geq 0, \quad \forall v, i \quad \text{(ex-post individual rationality)}.$$

A mechanism that satisfies the *dominant-strategy incentive compatibility (DSIC)* constraints is also called *truthful* or *strategyproof*. The DSIC constraints are sometimes weakened to *Bayesian incentive compatibility (BIC)*:

$$\mathbb{E}_{v_{-i}\sim P(\cdot|v_i)}\left[a(v_i, v_{-i}) \cdot v_i - p_i(v_i, v_{-i})\right] \geq \mathbb{E}_{v_{-i}\sim P(\cdot|v_i)}\left[a(b_i, v_{-i}) \cdot v_i - p_i(b_i, v_{-i})\right] \quad \forall v_i, b_i, i$$

In Bayesian mechanism design problems, it is common to consider the *interim mechanism* for each agent i defined by the average allocation and payment rules $\mathbb{E}_{v_{-i}\sim P(\cdot|v_i)}\left[a(v_i, v_{-i})\right]$ and $\mathbb{E}_{v_{-i}\sim P(\cdot|v_i)}\left[o_i(v_i, v_{-i})\right]$.

Remark 2.2 *Revenue-maximizing DSIC auction design.* An important special case of Definition 2.1 is revenue-maximizing DSIC auction design. Here, there are m items being sold to n bidders. The mechanism chooses either a deterministic assignment $o \in \{0,1\}^{n\times m}$ or a lottery over such assignments. In full generality, bidders may have types $v_i \in \mathbb{R}^{2^m}$ expressing their value for every bundle. In an important special case, *additive* bidders value each item individually, with their types represented by $v_i \in \mathbb{R}^m$. For additive bidders, the value of an assignment is the sum of the values for the assigned items. The auctioneer's performance goal is simply $\mathcal{L}(a,p,v) = \sum_i p_i(v_i, v_{-i})$, the total revenue.

Revenue-maximizing multi-bidder auction design is a natural problem to study and has immediate practical importance—many real-world auctions are run with the goal of maximizing revenue. Yet almost nothing analytical is known about optimal DSIC solutions to this problem, even in the seemingly simple case of additive valuations, beyond the single-item result of Myerson [1981], and the result of Yao [2017], which assumes valuation distributions with bivalued support.

This combination—little theoretical progress on a very natural and important problem—has meant that revenue-maximizing auction design has become a model problem for AMD. Since auction design has been such a major focus, we will focus on it heavily in this survey, and move somewhat freely between talk about agents (as in general mechanism design problems) and bidders (as in auction design). Of course, automated mechanism design can be used to design auctions with other goals, or to design other types of mechanisms entirely, and we will mention these other types of work where appropriate.

2.1 Characterizations of truthfulness

Here, we summarize some useful results about strategyproof mechanisms which tend to show up throughout the AMD literature.

2.1.1 *Convexity and truthfulness.* For agents with quasilinear utilities, there is a direct connection between truthfulness and convexity. The key connection is the concept of a cyclically monotone function. In convex analysis, a function is cyclically

monotone if and only if it is the (sub)gradient of some convex function (§24 of Rockafellar [1970]). It can also be shown [Rochet 1987] that the condition of cyclic monotonicity is equivalent to the DSIC constraints in Definition 2.1.[1]

The upshot of this for mechanism design is that every truthful mechanism induces a convex utility function for the mechanism participant, whose (sub)gradient is the allocation rule. Conversely, if and only if an allocation rule is cyclically monotone, it can be paired with a payment rule that will result in a truthful mechanism and a convex utility.

This relationship is defined for single-agent mechanisms. For multi-agent DSIC problems, it holds true for the mechanism faced by each bidder i holding v_{-i} fixed. For Bayesian problems, it holds true for each bidder's *interim* mechanism.

Many automated mechanism design methods ensure truthfulness by enforcing cyclic monotonicity of the allocation, or convexity of the utility.

2.1.2 *Convex conjugates and menu representations.* Recall that types v and outcomes o are dual to each other. Given a truthful mechanism with convex utility function $u(v)$, we can define the convex conjugate (§ 12 of Rockafellar [1970]) as $f^*(o) = \sup_{v \in \mathcal{V}} v \cdot o - f(v)$.

The conjugate has many nice properties (most importantly, it is always convex). In the mechanism design context, taking the conjugate $u^*(o)$ of a truthful mechanism's utility function has a natural interpretation as summarizing a *menu*: for each outcome o, $u^*(o)$ is the agent's payment, so the agent receives utility $o \cdot v_i - u^*(o)$ for the outcome. Equivalently, a truthful direct-revelation mechanism with allocation rule $a = \nabla u$ will pick the correct menu element on behalf of the agent.

2.1.3 *Affine maximizers and Roberts' theorem.* The above gives a very general characterization of truthful mechanisms. A specific class of truthful mechanisms of great importance is known as the class of *affine maximizers*. These can be seen as a generalized version of the VCG mechanism [Vickrey 1961; Clarke 1971; Groves 1973], and they inherit its nice properties.

Affine maximizers have a well-defined set of parameters that can be arbitrarily varied to optimize the mechanism design goal while always remaining within the constraints of Definition 2.1. This is why many AMD techniques restrict their search to affine maximizers.

What is lost by restricting to affine maximizers? In one sense, nothing is lost, if one must choose a mechanism that is strategyproof on an arbitrary type space. The reason is *Roberts' theorem* [Roberts 1979], which states that for a mechanism design problem on an unrestricted domain (so each agent could have any value for any outcome), with three or more outcomes, *all* DSIC mechanisms must be affine maximizers.

On the other hand, the unrestricted domain is not the right model for most mechanism design problems, so non-affine-maximizers may be truthful. For example, in an auction setting, the assumptions in Roberts' theorem would require the strange situation that bidders may have unbounded positive or negative values for receiving certain items, and moreover that the same holds for the items their opponents

[1]There is a particularly clear explanation of the connection between cyclic monotonicity and truthfulness in Börgers [2015], Chapter 5.

receive (i.e., there is the possibility of spite or altruism). Limiting mechanism design to affine maximizers therefore may mean giving up on finding the true optimal mechanism for realistic problems.

3. MATHEMATICAL OPTIMIZATION APPROACHES

The mechanism design problem (Definition 2.1) is a constrained optimization problem. In fact, if one considers a type distribution with finite support and known density, then it becomes a linear program: the objective is linear and each of the DSIC constraints is a linear constraint. This requires allowing for allocations in a continuous space; i.e., either divisible goods or randomized (so-called "lottery") allocations.

3.1 Linear programming as a computational tool

The papers that introduced the problem of automated mechanism design framed the problem in exactly this way [Conitzer and Sandholm 2003a; 2003b; 2004; Sandholm et al. 2007]. They transform the mechanism design problem into a linear program, and solve it using standard solvers. They also establish that deterministic AMD, without the randomized allocations that allow for linear programming, is NP-hard. With randomized allocations, linear programming remains a powerful tool, and work has continued using this basic approach on new problems [Guo and Conitzer 2010; Zhang and Conitzer 2021; Albert et al. 2015; Conitzer and Sandholm 2004].

3.1.1 *Discretizing the type space.* Linear programming approaches to AMD typically operate on an explicit description of a discrete type distribution, with decision variables indexed by each possible type or type profile, and translating mechanism design into linear programs poses some problems.

Large support of type distributions. It is common for type distributions to have an extremely large support. For combinatorial valuation functions (even assuming discrete possible values for each item), the number of possible types is doubly-exponential in the number of items. Even for very simple valuation structures such as additive valuations, if the type distribution has a continuous support, then approximating it on a grid will require a support whose size is exponential in the number of items. Moreover, even though the size grows "only" polynomially in the fineness of the grid, sufficiently-accurate grids even for small numbers of items create very difficult problem instances in practice [Dütting et al. 2024; Sandholm and Likhodedov 2015].

Approximation error due to discretization. Discretizing a continuous type distribution, or coarsening a discrete type distribution for tractability, unavoidably introduces error. The linear program outputs a mechanism that is defined only on the discrete or coarsened space. One can then consider rounding to the nearest discrete or coarse point to recover a mechanism in the original space, but this introduces violations of the incentive compatibility constraint. In the Bayesian mechanism design setting, there are recipes to transform approximate-BIC mechanisms into (exact) BIC mechanisms, while bounding the loss in revenue, welfare, etc. [Cai et al. 2012a; Cai et al. 2021; Conitzer et al. 2022; Daskalakis and Weinberg 2012]. Such techniques can be used to correct for the IC approximation that

is introduced by making use of discretization.

3.2 Mechanism design and duality

Of course, linear programs have duals, and thinking about duality has been theoretically fruitful in mechanism design. A line of several papers [Cai et al. 2012b; 2012a] lays out a reduction from multi-bidder *Bayesian* mechanism design to a single-bidder problem, culminating in a duality framework for finite-support type distributions [Cai et al. 2019]. Giannakopoulos and Koutsoupias [2018] also describe a duality framework for mechanism design and derive an auction format they call the *Straight-Jacket Auction*, which they can prove optimal for some cases.

Daskalakis et al. [2017], and some followup work [Kash and Frongillo 2016], formulate *single-bidder* mechanism design as dual to an optimal transport problem, which, as explicated by Kleiner and Manelli [2019], is equivalent to an infinite-dimensional linear programming problem. The primal problem is a search over a subset of convex functions corresponding to feasible mechanisms (as discussed briefly in Section 2.1.1). The structure of the dual provides useful information about optimal solutions, and in some cases dual solutions can certify optimality of mechanisms. This is theoretically useful and also motivates some of the differentiable economics approaches discussed in Section 5.

Kolesnikov et al. [2022] generalize this optimal transport approach to multi-bidder *Bayesian* mechanism design. Using their duality result combined with the idea of many-to-one bidder reduction and a bag of tricks from other works [Cai et al. 2012a; Alaei et al. 2019; Kleiner et al. 2021], they are able to formulate and explicitly solve a linear program to find an optimal interim auction for multiple bidders and multiple items. Many of the aforementioned papers reduced mechanism design to linear programming to prove theoretical results about tractability; Kolesnikov et al. [2022] uses many of those advances to numerically solve actual linear programs.

4. AUTOMATED MECHANISM DESIGN AND LEARNING THEORY

If one only has sample access to the type distribution, then the mechanism design problem (Definition 2.1) becomes a learning problem: the goal is to choose a function from some parameterized class (the class of feasible mechanisms) to optimize some loss (the mechanism design objective), hopefully generalizing from the training samples to the true distribution.

There are several advantages to this perspective. Requiring only samples from the distribution avoids the problems of large support mentioned in Section 3. Further, it is perhaps more natural to imagine that a mechanism designer has access to historical data from the population of bidders, rather than perfect knowledge of the population distribution. Moreover, there is already an extremely rich body of learning techniques and one can seek to apply these to mechanism design. In this section, we focus on a line of work motivated by "classical" machine learning and using techniques which come with strong generalization guarantees.

4.1 Single-parameter settings

Much work has focused on *single-parameter* (e.g., selling one item) auction design settings, where the Myerson auction is known to be optimal and strategyproof. The learning problem is to choose a function from the family of feasible Myerson

auctions. In the case of regular distributions, this just means choosing the Myerson reserve price; for more general distributions it may be more involved (e.g., Roughgarden and Schrijvers [2016]). This has been an enormously fruitful line of research and we cannot do it justice here, but we recommend the article of Guo et al. [2020] from a previous issue of SIGEcom Exchanges for a comprehensive overview.

4.2 Multi-parameter settings

In multi-parameter settings, there are currently no general characterizations of strategyproof mechanisms that lead to sample-efficient learning, but it is usually possible to restrict the learning problem to some nicer class of multi-parameter strategyproof mechanisms.

Learning simple auctions. A long line of work focuses on learning within classes of so-called *simple* auctions (e.g., combinations of Myerson auctions). It can sometimes be established that the best auction in some class of simple auctions gives a constant-factor approximation of the optimal revenue [Chawla et al. 2010; Hajiaghayi et al. 2007]. Moreover, it is sometimes possible to learn from samples efficiently over such classes [Morgenstern and Roughgarden 2016; 2015; Balcan et al. 2008; Feldman et al. 2014; Hsu et al. 2016].

Learning affine maximizers. As mentioned above, for completely general type spaces, only affine maximizers are strategyproof due to Roberts' theorem [Roberts 1979], and moreover affine maximizers form a parameterized class of auctions in a very convenient way. This motivates work on learning affine maximizers. The general idea is that functions—such as the revenue function—induced by affine maximizers have very nice structural properties that allow bounding their pseudodimension or Rademacher complexity—complexity measures in learning theory that can be used to bound the generalization error of learning from training samples. Sample complexity results are even better for certain restricted yet interesting subclasses of affine maximizers. The techniques can also be applied to classes of simple auctions [Balcan et al. 2016; 2018; Balcan et al. 2018].

5. DIFFERENTIABLE ECONOMICS

Differentiable economics uses tools from modern deep learning to solve problems in mechanism design. Like the aforementioned works in Section 4, it frames the mechanism design problem as a learning problem over samples from the type space; however, it focuses more on practical performance and flexibility rather than strictly on computational complexity and generalization guarantees. While some works do present sample complexity results [Dütting et al. 2024; Kuo et al. 2020], the main goal of differentiable economics is to use data-driven optimization to explore mechanism design in practice, assuming sufficient sample access for training complex mechanisms effectively.

At the broadest level, differentiable economics refers to using a gradient-based search process to optimize within some parameterized, differentiable class of mechanisms. Why might one want to take this approach to automated mechanism design? Modern function approximators for deep learning are both flexible and composable. If part of a problem has poorly-understood structure, one can simply use a universal

neural network of some kind; on the other hand, if information on special structure is available, or if there are particular problem constraints, this information can often be baked into the architecture. Due to the use of end-to-end gradient-based optimization, it is easy to freely combine and compose different architectural components.

The computational tools— first-order optimizers, autograd libraries such as TensorFlow [Abadi et al. 2015], Jax [Bradbury et al. 2018], and PyTorch [Ansel et al. 2024], and so on—are well-maintained and pleasant to use due to the enormous resources poured into their development by tech companies. And, as has been a surprise to the larger research community, even though in principle they are non-convex and should be intractable, solving deep learning problems using stochastic first-order methods just turns out to work really well, and this is equally true when the application happens to be mechanism design.

5.1 Introducing differentiable economics

Dütting et al. [2024] tackled the problem of optimal DSIC auction design and introduced the term "differentiable economics." One can extract a core recipe which remains useful for a very wide range of mechanism design problems (Definition 2.1):

(1) Use a flexible neural network to model the allocation (and payment) rules of the mechanism a, p as a function of type profiles; train the network on samples $v = (v_1, \ldots, v_i, \ldots, v_n) \sim P$ from the type distribution to maximize the performance goal $\mathcal{L}$ (e.g., revenue).
(2) Wherever possible, enforce problem constraints by careful design of the network architecture. In the world of deep learning, this is known as imposing an *inductive bias*.
(3) For other constraints where this is not possible, search for constraint violations (often by optimizing the network inputs) and use these violations to calculate a penalty term in the loss function.

Given that a direct-revelation mechanism is a function taking type profiles as inputs and outputting an outcome, differentiable function approximators can be used to represent mechanisms for many problem settings, and their flexibility often means that natural constraints of the problem can be designed into the architecture. This is therefore a very general-purpose recipe.

Dütting et al. [2024] introduce multiple concrete methods for learning auctions. One, known as RegretNet, works for auctions with any number of items and bidders, and imposes less inductive bias. Another, known as RochetNet, specializes to the single-bidder setting and is therefore able to impose more inductive bias to enforce IC. (There is also a third, MyersonNet, for single-item auctions, which we do not focus on here.)

5.1.1 *Multi-bidder auctions: RegretNet.* RegretNet (Figure 1) uses feedforward neural networks to represent the allocation and payment rules.

The input is the type profile. The output of the allocation rule is a matrix, with activation functions that ensure no item will ever be over-sold. The output of the payment network is a number between zero and one expressing the *fraction* of their received welfare to recover from each player as a payment; this, the allocation,

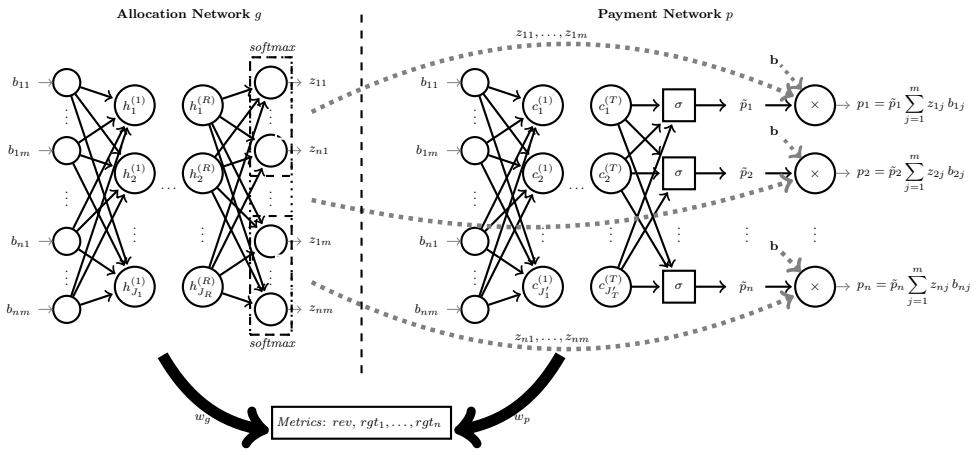


Fig. 1. A schematic of the RegretNet architecture, adapted from Dütting et al. [2024], for a setting with n buyers and m items. The neural network takes in bids as inputs and outputs both allocation and payment. Relevant metrics, such as regret and revenue, are then computed from these outputs and optimized by adjusting the network parameters through gradient descent.

and the types suffice to calculate the actual payment, which will never exceed a bidder's value for the items they get. These architectural design choices introduce an inductive bias, ensuring that individual rationality and item supply constraints are inherently satisfied.

DSIC, however, is not explicitly hardcoded into the architecture. Instead, it is enforced as a constraint during training, by penalizing *regret*: the expected maximum increase in utility that a bidder can achieve by misreporting their true value (in other words, the expected magnitude of the DSIC constraint violation). For a given truthful type profile, regret can be estimated by keeping the weights of the partially trained RegretNet fixed, and using gradient ascent on the bids themselves to maximize each bidder's utility from an untruthful report. Including the regret penalty in the training loss gradually drives the bidders' regret for truthful bidding to near zero.

The RegretNet approach is quite powerful and can be used to find high-performing mechanisms that are empirically almost perfectly strategyproof, and where there is often good reason to believe they are very close to the true optimal mechanisms. There is a problem, however. If an approximately DSIC mechanism is deployed, even a small (in terms of regret/utility) violation in DSIC may cause a larger shift in bidding behavior. Even worse, and as discussed in Curry et al. [2020], there may be very large (in terms of regret/utility) DSIC violations that only show up for a tiny proportion of types and may not be observed by the measurement methodology in Dütting et al. [2024]. Despite the empirical evidence that the mechanisms learned via differentiable economics tend to be near to truly strategyproof mechanisms, many mechanism designers—including those with both theoretical and practically-motivated interests—would prefer stronger guarantees.

5.1.2 *Exact DSIC for single bidders – RochetNet.* Fortunately, Dütting et al. [2024] also present an approach for a special case. *Single-bidder* DSIC mechanisms are well-characterized: as described in Section 2.1, they can be identified with convex, utility functions.

Dütting et al. [2024] use this characterization to design a single-bidder architecture that they call *RochetNet*, a schematic of which is shown in Figure 2. The utility function being convex (DSIC constraint), and non-decreasing and 1-Lipschitz (constrained to feasible allocations), is enforced at the architectural level.

Concretely, the learnable parameters of the network consist of a large number of menu elements, with each element specifying a bundle with the associated price. A null option with all-zero allocation and zero payment is added as a fixed menu element to ensure individual ratonality.

Given a bidder type as the input, the utility of each menu element can be evaluated, and a max operation is applied over all menu elements to select the highest-value item for the bidder, therefore ensuring exact IC. At training time, the non-differentiable max operation is approximated by a *softmax*. The menu elements, all being learnable parameters, are optimized by gradient descent. Because DSIC constraints are enforced via inductive bias in the architecture, the optimization proceeds in a straightforward way with no penalties or search for misreports.

Shen et al. [2019] give an architecture similar in practice to RochetNet, but they focus more on the conjugate interpretation (also discussed in Section 2.1). Dütting et al. [2024] and Shen et al. [2019] are both further able to build on the optimal transport formulation of mechanism design (Section 3.2) to formally prove the optimality of some mechanisms learned by their techniques.

5.2 Work related to Dütting et al. [2024]

RegretNet and RochetNet are two prototypical models which can help to think about related work — does a method enforce all constraints by inductive bias, like RochetNet, or does it have some penalty terms?

5.2.1 *Precursors to differentiable economics.* Conitzer and Sandholm [2007] design mechanisms by starting from a non-strategyproof mechanism, searching for counterexamples to strategyproofness, and repeatedly modifying the mechanism in response – similar in spirit to RegretNet. Narasimhan et al. [2016] and Dütting et al. [2015] use non-deep-learning ML techniques to find good mechanisms for problem settings with and without money. Sandholm and Likhodedov [2015] use a hill-climbing approach to optimize parameters of affine maximizer auctions.

5.2.2 *Further progress on auction design.* Some works have followed Dütting et al. [2024], focusing on the same auction problems and using essentially the same penalty-based training paradigm. Some involve improvements to the training algorithm [Rahme et al. 2021]. Others involve improvements to the architecture [Ivanov et al. 2022; Duan et al. 2022], both for the sake of regret/revenue performance and to allow for imposing additional inductive biases such as symmetry (which corresponds to bidder-anonymity).

Still others extend to auctions with new types of constraints: Feng et al. [2018] considers budget-constrained bidders, Tacchetti et al. [2022] optimizes for total participant welfare instead of revenue, Kuo et al. [2020] imposes a fairness constraint

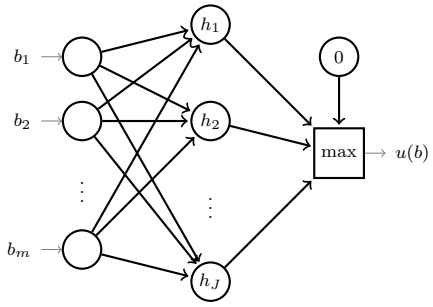


Fig. 2. A schematic of the RochetNet Architecture reproduced from Dütting et al. [2024]. RochetNet encodes the convex utility function associated with a truthful mechanism.

on the learned mechanisms (via training penalty), and Curry et al. [2022] extends to no-free-disposal settings where the number of items allocated to each bidder is enforced (via the architecture). Ravindranath et al. [2024] extends the RochetNet approach to sequential posted-menu mechanisms through the use of deep reinforcement learning.

5.2.3 *Problems beyond auctions.* Ravindranath et al. [2021] applies a RegretNet-like approach to two-sided matching, in particular to explore the space of mechanisms which trade off between the incompatible goals of stability and strategyproofness. Golowich et al. [2018] studies multi-facility location mechanism design, another mechanism design problem without payments. Ravindranath et al. [2023] studies the problem of data market design in economic theory, where the goal is to find a set of signaling schemes (statistical experiments) to maximize expected revenue to the information seller. Curry et al. [2024] extend the RochetNet approach to find optimal automated market makers, and extend the duality theory to prove them optimal. Wang et al. [2024] study contract design, and use specially-designed deep nets to model the utility function of the principal. A little further afield from these methods, Hossain et al. [2024] study the problem of multi-sender Bayesian persuasion.

5.2.4 *Affine maximizer auctions (again): a multi-bidder extension of RochetNet.* Affine maximizers are guaranteed to provide truthful, multi-bidder mechanisms, and as mentioned in Section 4, they have been used successfully for automated mechanism design and have well-understood learning-theoretic properties. They are also a good fit for a differentiable-economics inspired approach. In fact, if one considers the special case of an auction with one bidder, an affine maximizer is essentially just a menu, and in this sense this is one natural, multi-bidder generalization of RochetNet.

Curry et al. [2022] introduce this approach and find it works well. A particular advantage is that, by learning end-to-end, one can consider outcomes in the space of lotteries in addition to deterministic outcomes. A followup work takes advantage of the power of differentiable economics by adding attention layers everywhere, and finds increased performance [Duan et al. 2023]. Curry et al. [2024] further extend the approach to dynamic mechanism design problems, where multiple allocative decisions must be made over time.

Dütting et al. [2024] and Curry et al. [2022] observed that *overparameterization*

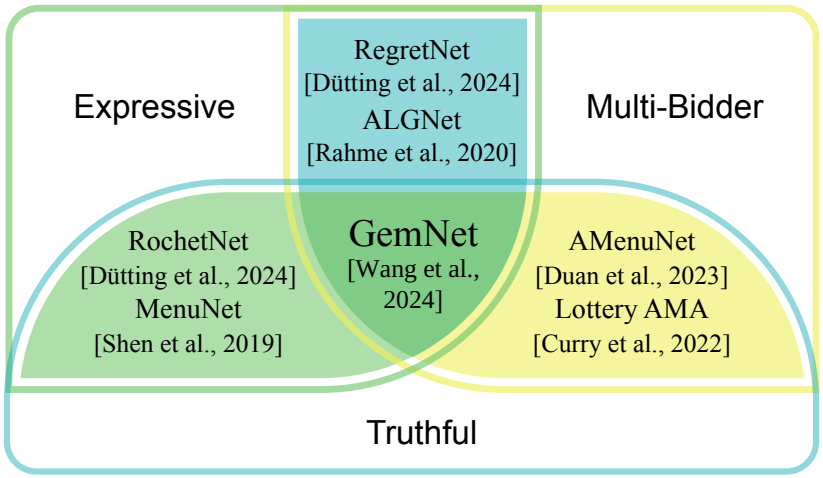


Fig. 3. The "trilemma" of current automated mechanism design reproduced from Wang et al. [2024]. Previously, no architecture had been exactly strategyproof, fully expressive, and multi-agent. GemNet is the first approach to achieve all three goals.

was useful during training of both RochetNet and affine maximizers, even though the learned mechanisms (in some cases, provably optimal) use very few menu elements by the end. By appealing to a concept from deep learning theory called *mode connectivity*, Hertrich et al. [2024] give a partial explanation for this phenomenon.

5.3 The current state of the art – GemNet

The aforementioned approaches demonstrate that perfectly strategyproof methods are achievable within differentiable economics. However, in the realm of multi-bidder auctions, which are particularly central to automated mechanism design, these methods must compromise either on full expressiveness or on exact strategyproofness. Wang et al. [2024] overcome this dilemma (see Figure 3) and give the current state-of-the-art method for differentiable economics: *GEneral Menu-based NETwork (GemNet)*.

GemNet extends the RochetNet architecture to learn a multi-bidder menu network that maps, for each bidder i, opponent bids v_{-i} onto a menu of allocations and payments for bidder i. As discussed in Section 2, as long as bidder i always chooses their favorite menu element, the mechanism will be strategyproof.

The problem is that if each bidder freely chooses their favorite menu element, the result might be that some items are over-sold. Some obvious ways present themselves to fix this (take into account other bidders' choices when choosing the menu, or adding a RegretNet-like activation to the allocation), but these turn out to destroy strategyproofness.

The main innovation in GemNet is to accommodate multi-bidder and multi-item settings by overcoming this problem, ensuring *menu compatibility*: even though bidders independently and freely choose the elements from their respective independent menus, the menus are designed such that their combined choices will not result in over-allocation of any item.

As a first step, the GemNet menu network is trained to optimize a combination of revenue and an *incompatibility loss*, which penalizes menus that could result in the over-allocation of items when bidders independently select their utility-maximizing menu elements. Although the incompatibility loss does not entirely eliminate menu incompatibility, it significantly reduces the likelihood of over-allocation.

Since any strategyproof mechanism can be represented by self-bid independent menus with bidder-optimization [Hammond 1979], for a network with sufficient capacity (i.e., enough hidden layers and nodes), GemNet's menu-based computational framework is without loss of generality, in that it can in principle learn the revenue-optimal auction with access to sufficient training data.

Exact menu compatibility in practice is achieved in a second step through a price adjustment algorithm involving mixed-integer linear programming (MILP), which fine-tunes menu prices post-training and pre-deployment to enforce compatibility across the entire continuous value domain. The approach leverages Lipschitz continuity of neural networks to extend compatibility from discrete grid points to continuous spaces, ensuring menu compatibility over the full domain and strategyproofness.

Empirically, GemNet demonstrates improved performance in various auction settings, outperforming existing methods such as AMAs in terms of revenue while achieving exact strategyproofness. Moreover, GemNet exhibits better interpretability through clear decision boundaries and agrees with known optimal solutions in specific scenarios. Figure 4 gives a nice illustration of this latter point.

Relationship with LP-based AMD methods. GemNet requires solving a series of mixed-integer linear programs (MILPs), which can be computationally demanding. Fortunately, because the trained networks typically exhibit only a minimal rate of over-allocation after training, it is often possible to preserve bidders' original choices in the trained menu (prior to any adjustments) for the majority of grid points. This strategy, among others, significantly reduces both the number of binary variables and the computational time needed to solve the MILPs (a decrease in running time of more than 99.99% in some settings). The size of the MILPs are substantially smaller than the size of the LPs that would be required in solving the same problem of AMD through earlier methods (as per Section 3). Moreover, although GemNet's price-adjustment method can in principle be adopted to a MILP-only framework for strategyproof mechanisms in continuous value domains, the use of deep learning is crucial in enabling local, "single-bidder focused" formulations.

6. CONCLUSION

6.1 Fruitful directions for future work

An even better characterization of strategyproof mechanisms. GemNet is fully expressive (i.e., it can represent any feasible mechanism, by appeal to universal approximation theorems for neural networks) and always exactly DSIC. But some parameter settings can represent infeasible mechanisms, hence the requirement of the post-training transformation. The post-training transformation is computationally costly, but designing an architecture that does not require it would probably require new theoretical breakthroughs in characterizing strategyproof mechanisms on restricted domains. With this improved understanding of feasible mechanisms,

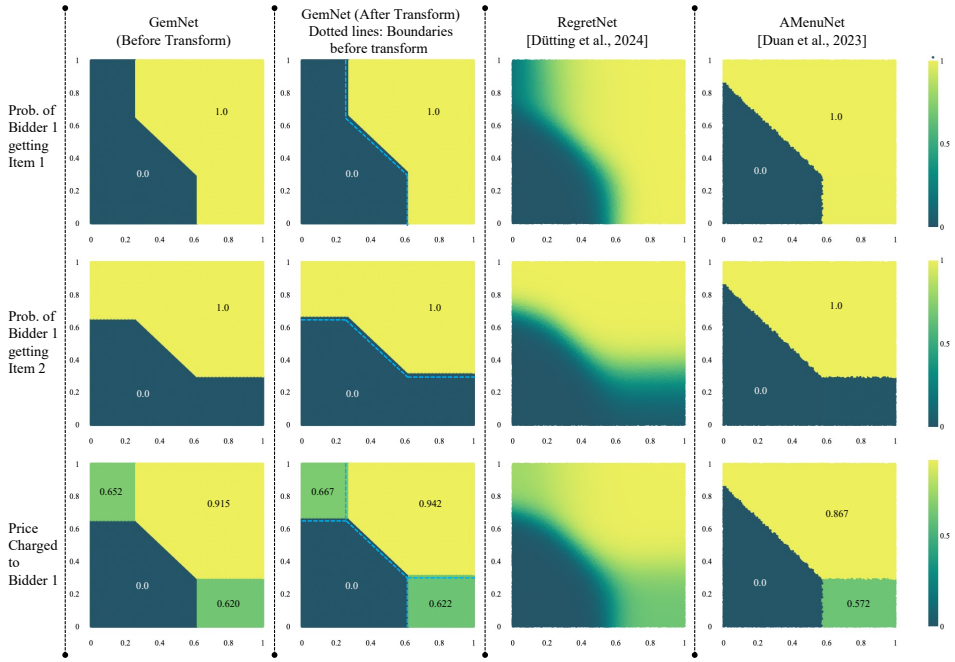


Fig. 4. This figure reproduced from Wang et al. [2024] shows the mechanism learned by several differentiable economics techniques — GemNet itself both before and after the price adjustments, RegretNet [Dütting et al. 2024], and the AMenuNet affine maximizer architecture [Duan et al. 2023]. The problem setting is selling 2 items to 2 additive agents with uniform type distributions on $[0, 1]$ — a problem setting for which the optimal auction is not known. The plots show the allocation rule for agent 1, holding agent 2's bid fixed. One can see that GemNet learns a mechanism with clear decision boundaries, while RegretNet appears to learn a smoother approximation of the same mechanism. The affine maximizer, meanwhile, learns a simpler mechanism, likely due to the reduced expressiveness of affine maximizers.

a differentiable economics pipeline could directly generate likely candidates for optimal and strategyproof mechanisms.

Sample-efficient differentiable economics. There are some generalization bounds provided in the work discussed in Section 5; e.g., in Dütting et al. [2024]. But those papers do not really treat sample efficiency as a major concern: most papers assume that it is easy to get at least hundreds of thousands of samples from the valuation distribution. This will not always be true, especially if these techniques are to be applied on real-world data.

Getting meaningful generalization bounds for deep learning is not easy; learning theory tends be too pessimistic. However, the architectures used for differentiable economics can be relatively small and can have special structure, so bridging the gap with the learning-theoretic work discussed in Section 4 could be possible and advantageous.

Automatically proving optimality. Differentiable economics has been used to find some optimal mechanisms [Dütting et al. 2024; Shen et al. 2019; Curry et al. 2024].

The idea is to search for a conjectured mechanism and then find a solution to a corresponding dual problem to certify optimality. Right now, these dual problems are solved by hand in an *ad hoc* way, which is both challenging and unreliable. An algorithm to attack the dual problem, paired with the existing techniques for searching through the primal space, could be extremely useful. And access to bounds from the dual during the optimization process could prove helpful for solving the primal problem more quickly.

Putting automated mechanism design to work for theorists. On the whole, these tools are not yet easy-to-use or reliable enough that a theoretical economist or mechanism designer would reach for them when studying a new problem. But progress on this front continues, and when automated mechanism design is truly mature, it should be a common part of the toolkit even for people who have no intrinsic interest in the computational techniques and would prefer to do as much as possible on a blackboard. One can imagine many scenarios: using automated mechanism design to generate conjectures, to explore the space of feasible solutions, to try to *falsify* conjectures, and much more. Progress here would involve both improvements in reliable training, easy-to-use software packages, and, crucially, visualization or other means of understanding learned mechanisms. There has been some partial progress on this front—moving from totally black-box neural networks to architectures that output interpretable menus—but the problem remains challenging, especially for mechanisms whose outputs live in higher-dimensional spaces that are intrinsically hard to visualize.

6.2 Wrapping up

There has been slow but steady progress on automated mechanism design in the decades since it was first introduced. The central idea shared by all automated mechanism design work is that mechanism design problems can be made to look a lot like other familiar optimization or learning problems from computer science, for which powerful computational tools already exist that can be brought to bear on mechanism design. The earliest work observed that mechanism design is a constrained optimization problem and so took an optimization-based approach. This is a very natural formulation and remains useful, but although solving LPs is efficient, the size of the LP formulations tend to scale badly (and require discrete type spaces). Other work has observed that mechanism design looks like a machine learning problem—choosing a function from some class, evaluating its performance on samples, and giving high-probability bounds on the estimation error. Most recently, work in *differentiable economics* brings to bear the techniques and pragmatic sensibilities of modern deep learning. This means giving up on some guarantees, but has also resulted in state-of-the-art results, including the discovery of new optimal mechanisms and interesting insights on very challenging problems.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Alaei, S., Fu, H., Haghpanah, N., Hartline, J., and Malekian, A. 2019. Efficient Computation of Optimal Auctions via Reduced Forms. *Mathematics of Operations Research 44,* 3 (Aug.), 1058–1086.

Albert, M., Conitzer, V., and Lopomo, G. 2015. Assessing the Robustness of Cremer-McLean with Automated Mechanism Design. *Proceedings of the AAAI Conference on Artificial Intelligence 29,* 1 (Feb.).

Alcobendas, M., Ji, J., Gokulakannan, H., Wami, D., Kapchits, B., Duteil, E. P., Satow, K., Roman, M. R. L., Diaz, O., Jr, A. A. D., and Kavoori, R. 2024. Optimizing Floors in First Price Auctions: An Empirical Study of Yahoo Advertising.

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., and Chintala, S. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Balcan, M.-F., Blum, A., Hartline, J. D., and Mansour, Y. 2008. Reducing mechanism design to algorithm design via machine learning. *Journal of Computer and System Sciences 74,* 8, 1245–1270.

Balcan, M.-F., Dick, T., and Vitercik, E. 2018. Dispersion for Data-Driven Algorithm Design, Online Learning, and Private Optimization. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. 603–614.

Balcan, M.-F., Sandholm, T., and Vitercik, E. 2018. A General Theory of Sample Complexity for Multi-Item Profit Maximization. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, Ithaca NY USA, 173–174.

Balcan, M.-F. F., Sandholm, T., and Vitercik, E. 2016. Sample complexity of automated mechanism design. *Advances in Neural Information Processing Systems 29*.

Börgers, T. 2015. *An introduction to the theory of mechanism design*. Oxford University Press, USA.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. 2018. JAX: composable transformations of Python+NumPy programs.

Cai, Y., Daskalakis, C., and Weinberg, S. M. 2012a. An algorithmic characterization of multi-dimensional mechanisms. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*. STOC '12. Association for Computing Machinery, New York, NY, USA, 459–478.

Cai, Y., Daskalakis, C., and Weinberg, S. M. 2012b. Optimal Multi-dimensional Mechanism Design: Reducing Revenue to Welfare Maximization. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. 130–139.

Cai, Y., Devanur, N. R., and Weinberg, S. M. 2019. A duality-based unified approach to bayesian mechanism design. *SIAM Journal on Computing 50,* 3, STOC16–160.

Cai, Y., Oikonomou, A., Velegkas, G., and Zhao, M. 2021. An efficient ϵ-bic to bic transformation and its application to black-box reduction in revenue maximization. In *Proceedings of*

the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA). Proceedings. Society for Industrial and Applied Mathematics, 1337–1356.

Chawla, S., Hartline, J. D., Malec, D. L., and Sivan, B. 2010. Multi-parameter mechanism design and sequential posted pricing. In *Proceedings of the forty-second ACM symposium on Theory of computing*. 311–320.

Clarke, E. H. 1971. Multipart pricing of public goods. *Public choice*, 17–33.

Conitzer, V., Feng, Z., Parkes, D. C., and Sodomka, E. 2022. Welfare-preserving ε-bic to bic transformation with negligible revenue loss. In *Web and Internet Economics*, M. Feldman, H. Fu, and I. Talgam-Cohen, Eds. Springer International Publishing, Cham, 76–94.

Conitzer, V. and Sandholm, T. 2003a. Automated mechanism design: Complexity results stemming from the single-agent setting. In *Proceedings of the 5th International Conference on Electronic Commerce*. ICEC '03. Association for Computing Machinery, New York, NY, USA, 17–24.

Conitzer, V. and Sandholm, T. 2003b. Automated mechanism design for a self-interested designer. In *Proceedings of the 4th ACM Conference on Electronic Commerce*. EC '03. Association for Computing Machinery, New York, NY, USA, 232–233.

Conitzer, V. and Sandholm, T. 2004. Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce*. EC '04. Association for Computing Machinery, New York, NY, USA, 132–141.

Conitzer, V. and Sandholm, T. 2007. Incremental Mechanism Design. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Curry, M., Sandholm, T., and Dickerson, J. 2022. Differentiable Economics for Randomized Affine Maximizer Auctions. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Curry, M., Thoma, V., Chakrabarti, D., McAleer, S., Kroer, C., Sandholm, T., He, N., and Seuken, S. 2024. Automated Design of Affine Maximizer Mechanisms in Dynamic Settings. *Proceedings of the AAAI Conference on Artificial Intelligence 38,* 9 (Mar.), 9626–9635.

Curry, M. J., Chiang, P.-Y., Goldstein, T., and Dickerson, J. 2020. Certifying Strategyproof Auction Networks. In *Neural Information Processing Systems*.

Curry, M. J., Fan, Z., and Parkes, D. C. 2024. Optimal automated market makers: Differentiable economics and strong duality. *arXiv preprint arXiv:2402.09129*.

Curry, M. J., Lyi, U., Goldstein, T., and Dickerson, J. P. 2022. Learning Revenue-Maximizing Auctions With Differentiable Matching. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 6062–6073.

Daskalakis, C., Deckelbaum, A., and Tzamos, C. 2017. Strong duality for a multiple-good monopolist. *Econometrica 85,* 3.

Daskalakis, C. and Weinberg, S. M. 2012. Symmetries and optimal multi-dimensional mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, Valencia Spain, 370–387.

Duan, Z., Sun, H., Chen, Y., and Deng, X. 2023. A Scalable Neural Network for DSIC Affine Maximizer Auction Design. *Advances in Neural Information Processing Systems 36*, 56169–56185.

Duan, Z., Tang, J., Yin, Y., Feng, Z., Yan, X., Zaheer, M., and Deng, X. 2022. A context-integrated transformer-based neural network for auction design. In *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. PMLR, 5609–5626.

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D. C., and Ravindranath, S. S. 2024. Optimal auctions through deep learning: Advances in differentiable economics. *Journal of the ACM 71,* 1, 1–53.

Dütting, P., Fischer, F., Jirapinyo, P., Lai, J. K., Lubin, B., and Parkes, D. C. 2015. Payment Rules through Discriminant-Based Classifiers. *ACM Trans. Econ. Comput. 3,* 1 (Mar.), 5:1–5:41.

Feldman, M., Gravin, N., and Lucier, B. 2014. Combinatorial auctions via posted prices. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 123–135.

Feng, Z., Narasimhan, H., and Parkes, D. C. 2018. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*. 354–362.

Frongillo, R. M. and Kash, I. A. 2021. General truthfulness characterizations via convex analysis. *Games and Economic Behavior 130*, 636–662.

Giannakopoulos, Y. and Koutsoupias, E. 2018. Duality and optimality of auctions for uniform distributions. *SIAM Journal on Computing 47,* 1, 121–165.

Golowich, N., Narasimhan, H., and Parkes, D. C. 2018. Deep learning for multi-facility location mechanism design. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 261–267.

Groves, T. 1973. Incentives in teams. *Econometrica: Journal of the Econometric Society*, 617–631.

Guo, C., Huang, Z., and Zhang, X. 2020. Sample complexity of single-parameter revenue maximization. *ACM SIGecom Exchanges 17,* 2 (Jan.), 62–70.

Guo, M. and Conitzer, V. 2010. Computationally Feasible Automated Mechanism Design: General Approach and Case Studies. *Proceedings of the AAAI Conference on Artificial Intelligence 24,* 1 (July), 1676–1679.

Hajiaghayi, M. T., Kleinberg, R., and Sandholm, T. 2007. Automated online mechanism design and prophet inequalities. In *AAAI*. Vol. 7. 58–65.

Hammond, P. J. 1979. Straightforward individual incentive compatibility in large economies. *The Review of Economic Studies 46,* 2, 263–282.

Hertrich, C., Tao, Y., and Végh, L. A. 2024. Mode connectivity in auction design. *Advances in Neural Information Processing Systems 36*.

Hossain, S., Wang, T., Lin, T., Chen, Y., Parkes, D. C., and Xu, H. 2024. Multi-sender persuasion–a computational perspective. *arXiv preprint arXiv:2402.04971*.

Hsu, J., Morgenstern, J., Rogers, R., Roth, A., and Vohra, R. 2016. Do prices coordinate markets? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 440–453.

Ivanov, D., Safiulin, I., Filippov, I., and Balabaeva, K. 2022. Optimal-er Auctions through Attention. *Advances in Neural Information Processing Systems 35*, 34734–34747.

Kash, I. A. and Frongillo, R. M. 2016. Optimal auctions with restricted allocations. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. 215–232.

Kleiner, A. and Manelli, A. 2019. Strong Duality in Monopoly Pricing. *Econometrica 87,* 4, 1391–1396.

Kleiner, A., Moldovanu, B., and Strack, P. 2021. Extreme points and majorization: Economic applications. *Econometrica 89,* 4, 1557–1593.

Kolesnikov, A. V., Sandomirskiy, F., Tsyvinski, A., and Zimin, A. P. 2022. Beckmann's approach to multi-item multi-bidder auctions. arXiv:2203.06837.

Kuo, K., Ostuni, A., Horishny, E., Curry, M. J., Dooley, S., Chiang, P.-y., Goldstein, T., and Dickerson, J. P. 2020. Proportionnet: Balancing fairness and revenue for auction design with deep learning. *arXiv preprint arXiv:2010.06398*.

Morgenstern, J. and Roughgarden, T. 2016. Learning Simple Auctions. In *Conference on Learning Theory*. PMLR, 1298–1318.

Morgenstern, J. H. and Roughgarden, T. 2015. On the pseudo-dimension of nearly optimal auctions. In *Advances in Neural Information Processing Systems*.

Myerson, R. B. 1981. Optimal Auction Design. *Mathematics of Operations Research 6,* 1 (Feb.).

Narasimhan, H., Agarwal, S. B., and Parkes, D. C. 2016. Automated mechanism design without money via machine learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

Rahme, J., Jelassi, S., and Weinberg, S. M. 2021. Auction learning as a two-player game. In *International Conference on Learning Representations*.

Ravindranath, S. S., Feng, Z., Li, S., Ma, J., Kominers, S. D., and Parkes, D. C. 2021. Deep Learning for Two-Sided Matching. arXiv:2107.03427.

Ravindranath, S. S., Feng, Z., Wang, D., Zaheer, M., Mehta, A., and Parkes, D. C. 2024. Deep reinforcement learning for sequential combinatorial auctions. arXiv:2407.08022.

Ravindranath, S. S., Jiang, Y., and Parkes, D. C. 2023. Data market design through deep learning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. Vol. 36. Curran Associates, Inc., 6662–6689.

Roberts, K. 1979. The characterization of implementable choice rules. *Aggregation and revelation of preferences 12,* 2, 321–348.

Rochet, J.-C. 1987. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of Mathematical Economics 16,* 2 (Jan.), 191–200.

Rockafellar, R. T. 1970. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J.

Roughgarden, T. and Schrijvers, O. 2016. Ironing in the Dark. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. EC '16. Association for Computing Machinery, New York, NY, USA, 1–18.

Sandholm, T. and Likhodedov, A. 2015. Automated Design of Revenue-Maximizing Combinatorial Auctions. *Operations Research 63,* 5 (Oct.), 1000–1025.

Sandholm, T. W., Conitzer, V., and Boutilier, C. 2007. Automated Design of Multistage Mechanisms. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Shen, W., Tang, P., and Zuo, S. 2019. Automated Mechanism Design via Neural Networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*.

Tacchetti, A., Strouse, D., Garnelo, M., Graepel, T., and Bachrach, Y. 2022. Learning truthful, efficient, and welfare maximizing auction rules. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*.

Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance 16,* 1, 8–37.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature 620,* 7972, 47–60.

Wang, T., Duetting, P., Ivanov, D., Talgam-Cohen, I., and Parkes, D. C. 2024. Deep contract design via discontinuous networks. *Advances in Neural Information Processing Systems 36*.

Wang, T., Jiang, Y., and Parkes, D. C. 2024. GemNet: Menu-Based, Strategy-Proof Multi-Bidder Auctions Through Deep Learning. In *Proceedings of the 2024 ACM Conference on Economics and Computation*.

Yao, A. C.-C. 2017. Dominant-Strategy versus Bayesian Multi-item Auctions: Maximum Revenue Determination and Comparison. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, Cambridge Massachusetts USA, 3–20.

Zhang, H. and Conitzer, V. 2021. Automated Dynamic Mechanism Design. In *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 27785–27797.

Randomized Apportionment Methods for Exact Proportional Representation: a Short Survey

HARIS AZIZ
The University of New South Wales

Apportionment is the problem of allocating seats to political parties in a parliament in proportion to their deserved representation or to allocate the number of representatives to states in proportion to their size. Throughout history, most of the focus is on deterministic methods to apportion seats among the groups which may favour bigger or smaller parties or may have some inherent mathematical limitations. We survey various randomized rules that achieve exact apportionment in expectation.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems; I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Economics*

General Terms: Theory, Algorithms, Economics

Additional Key Words and Phrases: Apportionment, Fair allocation, Randomisation, Social Choice

1. INTRODUCTION

Suppose there are three towns of sizes 30k, 80k and 90k from where a total of two representatives will be selected. Should the town with the largest population get two representatives? Or should the two most populous towns get one representative each? Or should the least populous town also have some chance of getting a representative? These types of questions concern what is called *apportionment*.

In apportionment, n disjoint groups are to be allocated a given number of slots k in proportion to the group sizes. The problem is ubiquitous in settings such as proportional representation of seats in the US congress, European Parliament, and the German Bundestag as well as various other committee selection settings [Balinski and Young, 1980, Birkhoff, 1976, Mayberry, 1978, Niemeyer and Niemeyer, 2008]. It has been studied in political science, economics, operations research and computer science. Apportionment is one of the most well-studied problems in social choice and political science with various books written on the topic [Balinski and Young, 1982, Pukelsheim, 2014, Young, 1994]. The classic book of Balinski and Young [1982] discusses history of apportionment rules used in the American Congress.

> *"This surprisingly difficult problem has concerned statesmen, political analysts and mathematicians for over two hundred years. The reason is the central importance of that apportionments plays in representative government. The difference of just one seat can be crucial in tipping the balance of power in a legislature. Hence the design of apportionment formulas is of abiding interest to politicians."*

Authors' addresses: haris.aziz@unsw.edu.au

Balinski and Young add that the problem is applicable to many other scenarios:

> *"Similar problems arise in many other settings, however. Teachers are assigned to courses in proportion to the number of students who register for them. Medical personnel are assigned to army units in proportion to the number of soldiers in each unit. Computers and support staff are allocated to divisions in a firm according to the measures of need and demand."*

In the settings mentioned above, each party (also called abstractly as group) i has relative size x_i with $\sum_{i=1}^{n} x_i = 1$. In many settings, the input is not the relative sizes but the actual group populations or group entitlements. Since the goal is proportional representation, each group requires kx_i slots which is termed as its *target quota* or *entitlement* q_i for group i:

$$q_i = kx_i.$$

Since $q_i = kx_i$ may not be an integer, we have to resort to *apportionment* which means that in order to allocate exactly k slots, some group may get slightly more or less that its target quota. We call the outcome $t = (t_1, \ldots, t_n)$ where $\sum_{i\in[n]} t_i = k$ where each t_i is an integer. A minimal requirement is that the outcome should satisfy *quota compliance*: each group should get quota that is the result of the target quota being rounded up or down: $t_i \in \{\lfloor q_i \rfloor, \lceil q_i \rceil\}$ for all $i \in [n]$.

Example 1 Apportionment Problem. *Suppose $q = (q_1, q_2, q_3) = (0.3, 0.8, 0.9)$. We wish to select $k = 2$ seats. One outcome is $t = (t_1, t_2, t_3) = (0, 1, 1)$ in which the last two entries are rounded up and the first entry is rounded down. An ex-post integral outcome is one in which two entries are rounded up and one entry is rounded down.*

Several apportionment procedures have been introduced in the literature such as the methods of Hamilton, Jefferson, Webster, Adams, and Hill [Balinski and Young, 1982]. Hamilton's method is essentially the method of the 'largest remainders' whereby each party is given its lower quota and then parties with the largest remainders are then given an extra seats to achieve the target seats. Many of the other methods such as Jefferson, Webster, Adams, and Hill are *divisor methods* in which each quota is divided by a given divisor such that when a rounding function is applied to each of the terms and then all the rounded terms are added, we get the desired number of seats. Different rounding functions correspond to different divisor methods. Most of the apportionment method discussed in theory and applied in practice are deterministic methods and each of them has some drawbacks. For example some method favours the bigger groups whereas some other favours the smaller groups. One particular disadvantage is not a design flaw of the method but is a consequence of a fundamental mathematical impossibility concerning deterministic methods. Balinski and Young [1982] proved that *no* deterministic apportionment procedure can simultaneously satisfy the following two of axioms: (1) *quota compliance* and (2) *population monotonicity* (if the ratio between the entitlements of two states i and j increases then it should be the case that the number of seats of i increases and the the number of seats of j decreases). However, if we use randomization, we can achieve the target quotas exactly (in expectation) which

means we can satisfy the two axioms. Randomization also has another benefit. It can ensure that every group has at least some probability of being represented. For example, in the three town example, we can ensure that each town has some probability of having a selected representative.

Note that any randomized method that achieves in expectation the expected target quota for each group also satisfies the following properties in expectation: (1) *Bias Condition*: The apportionment method should be free of bias, that is, it should neither favour large nor small parties' seats (2) *Independence Condition* (the number of seats assigned to the party depends only on its exact quota but not on the distribution of the quotas of other parties); (3) *House monotonicity* (if k increases then the seats of no party decreases).

How can we use randomization to achieve exact proportional representation in expectation and also achieve quota compliance ex post? We take a tour and visit some of the randomized apportionment methods that have been proposed in the literature.

2. A TOUR OF SOME RANDOMIZED APPORTIONMENT METHODS

We discuss some of the randomized apportionment methods.

2.1 Grimmet's Method

Grimmet [2004] presented a randomized apportionment rule that achieves the quota requirements. Grimmet's algorithm requires two successive randomized decisions, and one of them involves a continuous variable. The method satisfies the quota rule as do all the other randomized rules that we will discuss.

Grimmet's Method

(1) Chooses a permutation of the groups uniformly at random.
(2) Draw U uniformly at random from $[0,1]$, and let $Q_i = U + \sum_{j=1}^{i} q_j$. Allocate to each group i one seat for each integer contained in the interval $[Q_{i-1}, Q_i)$.

Note that $\sum_{j=1}^{i} q_i = k$. So the number of integers between U and $\sum_{j=1}^{i} q_i + U$ is k as well. Next, we observe that a particular group j's allocation is the number of integers between $U + \sum_{j=1}^{i-1} q_j$ and $U + \sum_{j=1}^{i-1} q_j + q_i$. This is either $\lfloor q_i \rfloor$ or $\lceil q_i \rceil$ so quota compliance is satisfied.

Example 2 Grimmet's Method. $q = (0.3, 0.8, 0.9)$ *We wish to select* $k = 2$ *additional seats. We choose a permutation uniformly at random and suppose it is* 123. *We draw* U *uniformly at random from* $[0,1]$ *and suppose it is* 0.2. *Then* $Q_0 = 0.2$, $Q_1 = 0.5$, $Q_2 = 1.3$, $Q_3 = 2.2$. *Then, the ex post outcome is* $r = (0,1,1)$ *that gives one seat each to the second and third party. The outcome can be different depending on what is the drawn value of* U.

2.2 Pipage Method

We describe the randomized method called Pipage that has proposed in mathematics and computer science [Gandhi, Khuller, Parthasarathy, and Srinivasan, 2006,

Deville and Tille, 1998, Ageev and Sviridenko, 2004]. The method can be viewed as starting from the root node of a tree and iteratively taking one of the two possible branches with particular probabilities. As we traverse down the tree, the number of fractional entries decreases by one in each step. In at most n steps, we reach a fully integral rounded vector. The probability with which various integral vectors are generated gives in expectation the original vector $(q_i)_{i\in[n]}$.

Pipage Method

Let $q^0 = q$. We iteratively and randomly modify q^0 in rounds. Denote $q^t = (q_1^t, q_2^t, ..., q_m^t)$ as the values at round t. In each round, we update the values of at most two indices while keeping the values of all other indices constant. Let $F^t = \{i \in C \mid q_i^t \in (0,1)\}$ be the set of indices that are fractional in round t. The update rule depends on the cardinality of F^t.
While $|F^t| \geq 2$, we arbitrarily select two indices $i, j \in F^t$ and run the following randomized update rule:

$$(q_i^{t+1}, q_j^{t+1}) = \begin{cases} \left(q_i^t + a, q_j^t - a\right) & \text{with probability } \frac{b}{a+b} \\ \left(q_i^t - b, q_j^t + b\right) & \text{with probability } \frac{a}{a+b} \end{cases}$$

where

$$a = \min\{c > 0 \mid q_i^t + c = 1 \text{ or } q_j^t - c = 0\}$$

and

$$b = \min\{c > 0 \mid q_i^t - c = 0 \text{ or } q_j^t + c = 1\}.$$

For all other indices $\ell \in C \setminus \{i, j\}$, we set $q_\ell^{t+1} = q_\ell^t$.
If $|F^t| < 2$, we terminate the algorithm and set $P_i = q_i^t$ for all $i \in C$.

Let us explain a few salient features of the method. At every step, when two entries are updated, the total amount in the vector remains the same: if one entry is decreased, the other entry is increased by the same amount. Now if one entry x is increased by a and decreased by b in the other branch, then it is increased by a with probability $b/(a+b)$ and decreased by b with probability $a/(a+b)$. Hence the expected net change is $ab/(a+b) - ab/(a+b) = 0$ as desired.

Example 3 Pipage Method. *Let us illustrate the Pipage method.* $q = (0.3, 0.8, 0.9)$. *We wish to select* $k = 2$ *seats. Then, the method works as follows in Figure 2.2. The probability of each outcome is equal to the product of probabilities along the path from the root node. For example, the probability of outcome* $(1, 1, 0)$ *is* $\frac{7}{9} \times \frac{1}{10} + \frac{2}{9} \times \frac{1}{10} = \frac{9}{90} = \frac{1}{10}$.

It is clear the Pipage method does not construct an explicit probability distribution over a polynomial number of integral allocations. In each step, it branches in two directions and possibly has an exponential number of integral allocations in its support. However a single polynomial-sized path of the decision tree is traversed to generate an integral allocation.

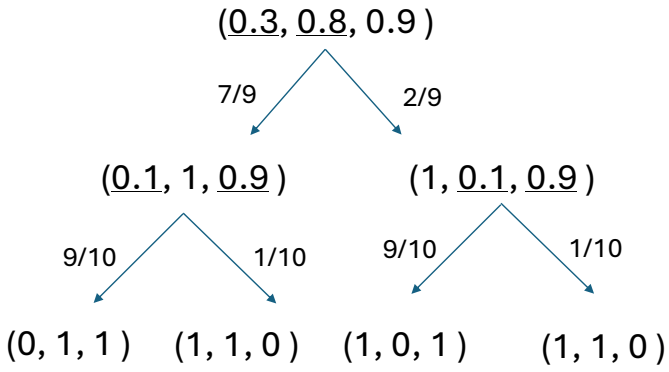


Fig. 1. Pipage Method. At each node except the leaf nodes, the underlined indices are selected and then a randomized update rule is run on them.

2.3 Careful Sliding Method

Next, we present the Careful Sliding Method. It has been presented as a subroutine for a 'peer selection' mechanism [Aziz, Lev, Mattei, Rosenchein, and Walsh, 2019] but it constitutes a simple and easy to apply rule for randomized apportionment. One feature of the method is that it returns an explicit probability distribution over integral apportionment and the size of the distribution is linear in the input size. The algorithm starts by rounding up the lowest quotas to obtain an integral allocation. The probability assigned to such an allocation is maximized subject to the condition that there is enough probability with which the higher quotas can also be rounded up to achieved their target quota in expectation. Once a given group has already achieved its expected quota, it is made *tight* and only the next higher groups are considered for rounding up. In the other direction, if there is a group with a high target quota and it needs to be rounded up in all subsequent integral allocations, it is made tight as well. In each step, a low quota or a high quota group is made tight. A low quota that is made tight is *not* rounded up in any subsequently generated integral allocations. A high quota that is made tight is rounded up in *all* subsequently generated integral allocations.

Before formally describing the method, we first present a simple illustrative example for how it works.

Example 4. *Suppose* $q = (0.3, 0.8, 0.9)$. *We wish to select* $k = 2$ *seats. The outcome of the Careful Sliding Method is as follows. The integral allocations and their corresponding probabilities are listed.*

$$S_1 = (1, 1, 0) : 0.1$$
$$S_2 = (1, 0, 1) : 0.2$$
$$S_3 = (0, 1, 1) : 0.7$$

The lottery computed indicates that the probability of outcome $(1, 1, 0)$ *is* 0.1*; the probability of outcome* $(1, 0, 1)$ *is* 0.2*; and the probability of outcome* $(0, 1, 1)$ *is* 0.7.

In the first step, we note that since $0.9 < 1$*, it does not need to be rounded up in the current and all subsequently generated integral outcomes. Hence, it is not labeled tight and will not be rounded up in the currently generated integral outcome. So we round the first two non-tight numbers* 0.3 *and* 0.8 *to get* $S_1 = (1, 1, 0)$*. The maximum probability that can be assigned to* $(1, 1, 0)$ *is* $1 - 0.9 = 0.1$ *Any higher probability makes it impossible to achieve* 0.9 *in expectation for the last entry. After* S_1 *and its corresponding probability are generated, we know that* 0.9 *will always be rounded up in* $S_2, S_3, \ldots$*, so it is labeled tight. So* $S_2 = (1, 0, 1)$ *is assigned probability* 0.2 *as giving any more probability does not leave enough probability for the second party to achieve its expected quota of* 0.8*. After this,* 0.3 *is labeled tight and the only entry that is not tight is* 0.8*. Finally,* $S_3 = (0, 1, 1)$ *with corresponding probability* 0.7*.*

Next, we present the method's description.

Careful Sliding Method

(1) Allocate to each group i, its lower quota $\lfloor q_i \rfloor$ to reduce the problem to that of allocating $\alpha = k - \sum_{i \in N} \lfloor q_i \rfloor$ seats with target $s_i = q_i - \lfloor q_i \rfloor$ for each group.

(2) **Initialize:** Label s_i for all i as *non-tight*. Set unallocated probability r to be 1. Set $\ell = 0$.
(we want to gradually relabel all s_i's as tight. We will iteratively generate a new integral outcome $S_{\ell+1}$ and its corresponding probability $p_{\ell+1}$ to achieve a partial lottery $[(S_1 : p_1), (S_2 : p_2) \ldots, (S_{\ell+1} : p_{\ell+1})]$.)

(3) Check if for highest non-tight number (say s_j), it is the case that the expected value of s_j can be obtained if we round up s_j in all $S_{\ell+1}, \ldots$. If yes, we decrement α by one and set s_j to be *tight*. The entry s_j will be rounded up in allocations $S_{\ell+1}, \ldots$. Repeat until $\alpha = 0$ or the condition does not hold.

(4) Increment ℓ by one. Round up the smallest non-tight α numbers to get S_ℓ. The corresponding probability p_ℓ of S_ℓ is the maximum feasible probability such that it still allows enough unallocated probability $r - p_\ell$ to achieve the highest non-tight number. Add p_ℓ and the corresponding rounded outcome S_ℓ to the distribution. Set r to $r - p_\ell$. If some s_i is already achieved in partial lottery $[(S_1 : p_1), \ldots, (S_\ell : p_\ell)]$, then it cannot be rounded up in future allocations so, its relabelled as tight (will be rounded down in all next allocations $S_{\ell+1}, \ldots$).

(5) Repeat (3), (4) until $\alpha = 0$.

(6) Return distribution $[(S_1 : p_1), \ldots, (S_\ell : p_\ell)]$.

2.4 Sampford's Method

Finally, we describe a simple method due to Sampford [1967] that has recently been shown to satisfy desirable axiomatic properties. The method first allocates each group i, its lower quota $\lfloor q_i \rfloor$ to reduce the problem to that of allocating $\alpha = k - \sum_{i \in N} \lfloor q_i \rfloor$ seats with target $p_i = q_i - \lfloor q_i \rfloor$ for each group. The first party

to be given an extra seat is selected with probability proportional to p_i. For all the subsequent draws, parties are drawn with probabilities proportional to $\frac{p_i}{1-p_i}$. If any of the α groups is repeated, we restart to sample α groups.

Sampford's Method

(1) Already allocate each group i, its lower quota $\lfloor q_i \rfloor$ to reduce the problem to that of allocating $\alpha = k - \sum_{i \in N} \lfloor q_i \rfloor$ seats with target $p_i = q_i - \lfloor q_i \rfloor$ for each group.
(2) Select α groups with replacement, the first drawing being made with probabilities p_i for each group i and all the subsequent ones with probabilities proportional to $\frac{p_i}{1-p_i}$.
(3) **If** the α drawn parties are distinct, give one additional seat to each of them;
Else, start over.

One important issue is that the method may not terminate as we could repeatedly get a party that gets two of the α seat which violates quota compliance. However, the sampling can also be implemented in a polynomial-time manner [Grafström, 2009] although the polynomial-time algorithm's description is not as simple as the original method's. An alternative and simpler method than Sampford is *weighted random sampling* in which each of the α parties is selected with probability proportional to p_i. However such a naive way to sample does not respect that each party's seats are rounded up with probability p_i.

Example 5 Sampford's Method. $p = (0.3, 0.8, 0.9)$ *We wish to select* $k = 2$ *additional seats. We first select the first party with probabilities proportional to* $0.3, 0.8, 0.9$ *respectively. Suppose the first party selected is the third one. One of the parties from party 1, party 2, and party 3 are selected with probabilities proportional to* $0.3/0.7$*,* $0.8/0.2$*, and* $0.9/0.1$ *respectively. Equivalently, party 1 is selected with probability* $(3/7)/((3/7)+4+9)$*, party 2 is selected with probability* $4/((3/7)+4+9)$*, and party* 3 *is selected with probability* $9/((3/7)+4+9)$.

3. DISCUSSION AND RESEARCH DIRECTIONS

We have illustrated key randomized apportionment methods. The growing literature on randomized apportionment fits in the wider framework of "*best of both worlds fairness*" whereby the goal is to simultaneously achieve strong fairness properties in expectation and approximately fair properties ex-post [Aziz, Freeman, Shah, and Vaish, 2024].

Recently, there has been a surge in designing or identifying apportionment rules that satisfy further axiomatic properties [Gölz, Peters, and Procaccia, 2022, Correa, Gölz, Schmidt-Kraepelin, Tucker-Foltz, and Verdugo, 2024]. For example, Gölz et al. [2022] use the techniques of Gandhi et al. [2006] to additionally achieve a strong specification of *house monotonicity* (if the number of seats increases, no one's seats are reduced).

Although randomized apportionment achieves population monotonicity in expectation, when parties or voters care about joint events, such as whether a coalition

of parties reaches a majority, further care needs to be made when generating a random outcome. Correa et al. [2024] observe that when parties already get their lower quotas, the decision about apportioning the remainders reduces to *"probability proportional to size" sampling without replacement* that is well-studied in mathematical statistics [Brewer and Hanif, 2013] and has dozens of well-established methods. They focus on *threshold monotonicity* that requires that for any coalition of parties T, if the tarqet quota of each party in T weakly increases and each party outside T weakly decreases, then the random variable describing the total number of seats awarded to T first-order stochastically dominates the corresponding random variable before the reinforcement. Correa et al. [2024] showed that the Sampford's Method is especially suitable as it satisfies a weaker version of threshold monotonicity. On the other hand, the other methods in this exposition violate selection monotonicity. Correa et al. [2024] ask whether Sampford's Method or some other method satisfies threshold monotonicity.

The recent results indicate that although apportionment has been studied for centuries, there are still interesting ongoing developments to understand randomized apportionment methods and their axiomatic properties. Future work on randomized apportionment promise to have a similar trajectory as the well-established literature on deterministic apportionment rules [Balinski and Young, 1982, Pukelsheim, 2014, Young, 1994]. This will entail several important steps including

- expanding the toolkit of apportionment rules and classifying them into various families;
- formalizing axiomatic properties for rules;
- understanding which rules satisfy what properties; and
- understanding what subsets of axioms are compatible or impossible to satisfy simultaneously.

For the multitude of axioms considered in the literature on deterministic rules, there may be multiple ways to generalize them to the rich class of randomized rules. Finally, axiomatically characterizing desirable randomized apportionment rules with respect to various axiomatic properties remains a major open question.

4. ACKNOWLEDGMENTS

This work was supported by the NSF-CSIRO grant on "Fair Sequential Collective Decision-Making" (RG230833). Thanks to Riley Baird, Sreoshi Banerjee, Patrick Lederer, and Irene Lo for comments.

REFERENCES

A. A. Ageev and M. I. Sviridenko. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization*, 8(3):307–328, 2004. doi: 10/dv46g4.

H. Aziz, O. Lev, N. Mattei, J. S. Rosenchein, and T. Walsh. Strategyproof peer selection using randomization, partitioning, and apportionment. *Artificial Intelligence*, 275:295–309, 2019.

H. Aziz, R. Freeman, N. Shah, and R. Vaish. Best of both worlds: Ex ante and ex post fairness in resource allocation. *Operations Research*, pages 1317–1750, 2024.

M. Balinski and H. P. Young. *Fair Representation: Meeting the Ideal of One Man, One Vote*. Yale University Press, 1982. (2nd Edition [with identical pagination], Brookings Institution Press, 2001).

M. L. Balinski and H. P. Young. The webster method of apportionment. *Proceedings of the National Academy of Sciences (PNAS)*, 77(1):1–4, 1980.

G. Birkhoff. House monotone apportionment schemes. *Proceedings of the National Academy of Sciences (PNAS)*, 73(3):684–686, 1976.

K. RW Brewer and M. Hanif. *Sampling with unequal probabilities*, volume 15. Springer Science & Business Media, 2013.

J. Correa, P. Gölz, U. Schmidt-Kraepelin, J. Tucker-Foltz, and V. Verdugo. Monotone randomized apportionment. *CoRR*, abs/2405.03687, 2024.

J-C. Deville and Y Tille. Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101, 1998. doi: 10.1093/biomet/85.1.89.

R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM*, 53(3): 324–360, 2006.

P. Gölz, D. Peters, and A. D. Procaccia. In this apportionment lottery, the house always wins. In *The 23rd ACM Conference on Economics and Computation (ACM EC)*, page 562. ACM, 2022.

A. Grafström. Non-rejective implementations of the Sampford sampling design. *Journal of Statistical Planning and Inference*, 139(6):2111–2114, 2009.

G. Grimmet. Stochastic apportionment. *The American Mathematical Monthly*, 111 (4):299–307, 2004.

J. P. Mayberry. Quota methods for congressional apportionment are still non-unique. *Proceedings of the National Academy of Sciences (PNAS)*, 75(8):3537–3539, 1978.

H. F. Niemeyer and A. C. Niemeyer. Apportionment methods. *Mathematical Social Sciences*, 56:240–253, 2008.

F. Pukelsheim. *Proportional Representation: Apportionment Methods and Their Applications*. Springer, 2014.

Michael R. Sampford. On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54(3-4):499–513, 1967. doi: 10/ccb3kz.

H. P. Young. *Equity: in Theory and Practice*. Princeton University Press, 1994.

Liquid Democracy: An Annotated Reading List

GEORGIOS PAPASOTIROPOULOS
University of Warsaw, Poland
and
ULRIKE SCHMIDT-KRAEPELIN
TU Eindhoven, Netherlands

Abstract: Liquid democracy is a democratic paradigm that introduces new challenges for researchers in fields around collective decision-making and, hence, it prompts a variety of compelling questions well-suited to the EC community. In this overview, we present a selection of papers that capture the breadth of research directions in this area.

Categories and Subject Descriptors: [**Theory of computation**]: Theory and algorithms for application domains—*Algorithmic game theory and mechanism design*; [**Applied computing**]: Computing in government—*Voting / election technologies*

General Terms: Algorithms; Design; Economics

Additional Key Words and Phrases: Computational Social Choice, Voting, Elections, Liquid Democracy, Delegations

Liquid democracy is a hybrid voting model that bridges the gap between direct and representative democracy. Direct democracy is often seen as impractical due to the scale and complexity of many elections and the subjects under consideration, while representative democracy restricts voters' influence and participation in decision-making. In liquid democracy, voters have the flexibility to choose whether to cast their own votes if they feel informed about the issue at hand or to delegate their voting power to someone they believe is more knowledgeable on the matter. These delegations, in a transitive manner, allow those who decide to cast a ballot to do so with a voting power that represents both their own vote and the votes of everyone who has (directly or indirectly) entrusted their vote to them.

For practical purposes, liquid democracy is mostly intended for low-stake decision-making within mid-sized communities or organizations. From a theoretical viewpoint, liquid democracy opens a rich array of research questions and models that we believe could be of interest to the EC community. The literature can be roughly categorized along two dimensions: We organize the proposed models into four categories, including epistemic models, ranked-delegation models, specialized frameworks, and models capturing strategic behavior; orthogonally, the methodologies range from algorithmic, axiomatic, and game-theoretic to statistical questions, and empirical analyses – with some works fitting into multiple groups across both dimensions. Our goal is to highlight the field's breadth and its latest developments while presenting works that also offer a more comprehensive perspective by contex-

Authors' addresses: gpapasotiropoulos@gmail.com, u.schmidt.kraepelin@tue.nl

tualizing earlier research. As a result, the selected papers are often the most recent work in their respective research lines. We hope that our list serves as a useful starting point for interested readers, shedding light on some of the most promising areas for future research in liquid democracy. This list is neither exhaustive nor fully representative, as many interesting, closely related, and important—if not foundational—works had to be omitted due to space constraints.

(1) [Berinsky, Halpern, Halpern, Jadbabaie, Mossel, Procaccia, and Revel, 2025] — The paper "Tracking Truth with Liquid Democracy" examines the *epistemic model* of liquid democracy, which addresses whether liquid democracy is more effective than direct democracy at uncovering a ground truth in a binary decision setting. Previous studies on this model (e.g., Kahng et al. (2021) and Caragiannis et al. (2019)) suggested that power concentration in liquid democracy may significantly lower its accuracy compared to direct democracy, even when voters delegate only to those more competent in identifying the ground truth. In contrast, this paper identifies models under which liquid democracy surpasses direct democracy in accurately uncovering the ground truth.

(2) [Kavitha, Makino, Schlotter, and Yokoi, 2024] — Among more general results, the paper "Arborescences, Colorful Forests, and Popularity" presents a polynomial-time, combinatorial, primal-dual algorithm for the *popular arborescence problem*, a problem whose computational complexity had been open since the earlier work of Kavitha et al. (2020). This algorithm has applications in liquid democracy with ranked delegations—a model designed to address delegation cycles by allowing voters to specify a set of possible delegations along with a preference ranking over them. Given these preferences, an arborescence assigns delegations to voters, and a popular arborescence (if one exists) is one that is preferred by a majority of voters over any alternative arborescence.

(3) [Utke and Schmidt-Kraepelin, 2023] — The paper "Anonymous and Copy-Robust Delegations for Liquid Democracy" builds upon a model suggested by Brill et al. (2022), and studies axiomatic properties of delegation rules for liquid democracy with *ranked delegations* (as discussed in (2)). While the model of Brill et al. requires that the voting weight of each voter is assigned to exactly one other voter, this work relaxes this assumption and allows to (fractionally) distribute the voting weight over multiple representatives. The authors first present an axiomatic impossibility theorem in the setting of Brill et al. and then show that a fractional delegation rule suggested in the literature (Brill (2018)) resolves this impossibility and can be computed in polynomial time.

(4) [Tyrovolas, Constantinescu, and Elkind, 2024] — The paper "Unravelling Expressive Delegations: Complexity and Normative Analysis" studies a generalization of the ranked delegation model (see (2) and (3)) in which voters can submit delegations in form of boolean functions, thereby enhancing the expressivity of a ballot to allow for the communication of conditional preferences (e.g., based on the majority opinion of others). This model was first suggested by Colley et al. (2020) who presented several ways for *unravelling* the voters preferences into one valid ballot per voter. While focusing on binary decisions, the authors present, among other results, computational dichotomies for two natural unravelling approaches: a utilitarian and an egalitarian one.

(5) [Markakis and Papasotiropoulos, 2024] — The paper "As Time Goes By: Adding a Temporal Dimension to Resolve Delegations in Liquid Democracy" proposes a framework in which the decision-making moment is preceded by an extended deliberation period, allowing voters to declare delegation choices (in a manner similar to the one examined by Brill et al. (2022)) and revise them at each step in response to new information or changes in others' opinions. This process aims to identify suitable representatives for all voters in cases where delegations at the (final) moment of decision prove inadequate or infeasible, such as in scenarios where voters' delegations lead to delegation cycles. The authors introduce axioms and examine their compatibility with efficient algorithms, mainly drawing on techniques and results from temporal graph theory.

(6) [Colley and Grandi, 2022] — The paper "Preserving Consistency in Multi-Issue Liquid Democracy" falls into the realm of liquid democracy with *interdependent issues* (see, e.g., Christoff and Grossi (2017)), where the same set of voters decides upon a range of such issues, and for each one, voters decide whether to vote themselves or to delegate, which may lead to inconsistent, infeasible ballots. The authors unify the approaches from earlier works on a similar model by Jain et al. (2022) and Brill and Talmon (2018) and they show that resolving the inconsistencies by minimizing either the number of ignored delegations or the number of changes to the votes is computationally hard. In response, they suggest that voters submit priorities over the issues which can then be used to find consistent votes in polynomial time.

(7) [Köppe, Kouteckỳ, Sornat, and Talmon, 2024] — The paper "Fine-Grained Liquid Democracy for Cumulative Ballots" builds upon an idea of Brill and Talmon (2018) and proposes a model where voters can distribute a unit of support across various bundles of election options (e.g., bundles of projects in a Participatory Budgeting scenario), and, if desired, delegate the precise distribution within each bundle to other voters. Delegation cycles or conflicts – such as when voters allocate no support to options that voters they represent wish to fund – necessitate centralized methods for resolving delegations that satisfy specific axiomatic guarantees. By establishing a relation to Nash equilibria, the authors use mainly the fixed-point theory to study the existence, structure, and computability of satisfactory solutions.

(8) [Bloembergen, Grossi, and Lackner, 2019] — The paper "On Rational Delegations in Liquid Democracy" introduces a model of elections on a binary issue where (either deterministic or probabilistic) voters are represented as vertices on a graph, and they can choose to vote directly with a specific accuracy related to their preferred outcome at a certain effort, or delegate their vote to a neighbor at no cost to inherit that neighbor's accuracy; a model motivated by similar considerations as the one examined by Alouf-Heffetz et al. (2025). Theoretical results establish the existence of Nash equilibria and analyze quality with respect to utilitarian social welfare and average accuracy in certain classes of the game in which voters strive to balance the accuracy achieved with the effort expended. Experimental simulations on synthetic network topologies assess the performance of delegation compared to direct voting, the number and quality of voters acting as ultimate representatives, and the existence of cycles.

(9) [D'Angelo, Delfaraz, and Gilbert, 2022] — The paper "Computation and Bribery of Voting Power in Delegative Simple Games" studies a generalization of voting games, where voters, represented as (weighted) nodes in a social network, acquire power through the transitive delegation structure of liquid democracy, reflecting the relative influence of each. The authors present a pseudo-polynomial time algorithm for the computationally hard problems of calculating voters' power expressed via the Banzhaf and Shapley-Shubik indices in these cooperative games, originally introduced by Zhang and Grossi (2021). They then examine (from the perspective of approximation and parameterized algorithms) the bribery problem, which aims to identify which voters should be influenced and how, within a budget constraint, to change their delegations in order to maximize or minimize the voting power or the final voting weight of a specific voter; this is closely related to the control problem examined (among other problems) by Alouf-Heffetz et al. (2025) with a focus on computational complexity and differing objectives.

(10) [Kling, Kunegis, Hartmann, Strohmaier, and Staab, 2015] — The paper "Voting Behaviour and Power in Online Democracy: A Study of LiquidFeedback in Germany's Pirate Party" analyzes real-world data from one of the most prominent software systems for online voting, which allows users to delegate their votes to others with the flexibility to adjust their choices over time (as studied theoretically in the model of Markakis et al.). Focusing on the platform's largest installation, the study observes a 4-year period involving around 14k users casting hundreds of thousands of votes and tens of thousands of delegations, on 6.5k proposals. The authors focus on understanding (i) the dynamics of voting and delegation behavior and (ii) the assessment of power each voter holds in the system and how this power is being used.

Acknowledgements

Georgios Papasotiropoulos is supported by the European Union (ERC, PRO-DEMOCRATIC, 101076570). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Alouf-Heffetz, Shiri, Łukasz Janeczko, Grzegorz Lisowski, and Georgios Papasotiropoulos (2025). "The Cost Perspective of Liquid Democracy: Feasibility and Control". In: *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI'25)*.

Berinsky, Adam J, Daniel Halpern, Joseph Y Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D Procaccia, and Manon Revel (2025). "Tracking Truth with Liquid Democracy". In: *Management Science*. Forthcoming.

Bloembergen, Daan, Davide Grossi, and Martin Lackner (2019). "On Rational Delegations in Liquid Democracy". In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19)*, pp. 1796–1803.

Brill, Markus (2018). "Interactive democracy". In: *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'18)*, pp. 1183–1187.

Brill, Markus, Théo Delemazure, Anne-Marie George, Martin Lackner, and Ulrike Schmidt-Kraepelin (2022). "Liquid Democracy with Ranked Delegations". In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22)*, pp. 4884–4891.

Brill, Markus and Nimrod Talmon (2018). "Pairwise Liquid Democracy". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, pp. 137–143.

Caragiannis, Ioannis and Evi Micha (2019). "A Contribution to the Critique of Liquid Democracy." In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*, pp. 116–122.

Christoff, Zoé and Davide Grossi (2017). "Binary Voting with Delegable Proxy: An Analysis of Liquid Democracy". In: *Electronic Proceedings in Theoretical Computer Science* 251, pp. 134–150.

Colley, Rachael and Umberto Grandi (2022). "Preserving Consistency in Multi-Issue Liquid Democracy". In: *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*, pp. 201–207.

Colley, Rachael, Umberto Grandi, and Arianna Novaro (2020). "Smart Voting". In: *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20)*, pp. 1734–1740.

D'Angelo, Gianlorenzo, Esmaeil Delfaraz, and Hugo Gilbert (2022). "Computation and Bribery of Voting Power in Delegative Simple Games". In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS'22)*, pp. 336–344.

Jain, Pallavi, Krzysztof Sornat, and Nimrod Talmon (2022). "Preserving consistency for liquid knapsack voting". In: *Proceedings of the 19th European Conference on Multi-Agent Systems (EUMAS'22)*. Springer, pp. 221–238.

Kahng, Anson, Simon Mackenzie, and Ariel Procaccia (2021). "Liquid democracy: An algorithmic perspective". In: *Journal of Artificial Intelligence Research* 70, pp. 1223–1252.

Kavitha, Telikepalli, Tamás Király, Jannik Matuschke, Ildikó Schlotter, and Ulrike Schmidt-Kraepelin (2020). "Popular branchings and their dual certificates". In: *Proceedings of the 21st International Conference on Integer Programming and Combinatorial Optimization (IPCO'20)*. Springer, pp. 223–237.

Kavitha, Telikepalli, Kazuhisa Makino, Ildikó Schlotter, and Yu Yokoi (2024). "Arborescences, Colorful Forests, and Popularity". In: *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'24)*, pp. 3724–3746.

Kling, Christoph, Jérôme Kunegis, Heinrich Hartmann, Markus Strohmaier, and Steffen Staab (2015). "Voting Behaviour and Power in Online Democracy: A Study of Liquid-

Feedback in Germany's Pirate Party". In: *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM'15)*, pp. 208–217.

Köppe, Matthias, Martin Koutecký, Krzysztof Sornat, and Nimrod Talmon (2024). "Fine-Grained Liquid Democracy for Cumulative Ballots". In: *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS'24)*, pp. 1029–1037.

Markakis, Evangelos and Georgios Papasotiropoulos (2024). "As Time Goes By: Adding a Temporal Dimension to Resolve Delegations in Liquid Democracy". In: *Proceedings of the 8th International Conference on Algorithmic Decision Theory (ADT'24)*, pp. 48–63.

Tyrovolas, Giannis, Andrei Constantinescu, and Edith Elkind (2024). "Unravelling Expressive Delegations: Complexity and Normative Analysis". In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI'24)*, pp. 9918–9925.

Utke, Markus and Ulrike Schmidt-Kraepelin (2023). "Anonymous and Copy-Robust Delegations for Liquid Democracy". In: *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS'23)*.

Zhang, Yuzhe and Davide Grossi (2021). "Power in Liquid Democracy". In: *Proceedings of the 35th AAAI conference on Artificial Intelligence (AAAI'21)*, pp. 5822–5830.