Table of Contents

ACM SIGecom Exchanges, Vol. 23, No. 1, July 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

We are pleased to present another exciting issue of the SIGecom Exchanges. The Exchanges seeks to keep the EconCS community abreast of relevant news and exciting research directions. This Summer 2025 issue has some of both. On the news side, we have a letter from the SIGecom executive committee, summarizing ongoing and upcoming SIG activities, and a recap of the 2025 Winter Meeting on Generative AI and Market Design. We also have three letters, a research survey, and an annotated reading list, previewed below.

The first of our three letters, by Nikhil Devanur, Renato Paes Leme, and Okke Schrijvers, discuss some of the common pitfalls they've seen among those transitioning from academic work in EconCS to work in industry. They highlight some of the most helpful skills to pick up in preparation, e.g. in the domains of software engineering and machine learning, and suggest resources for those looking to improve and hit the ground running.

Our second letter, from Pranav Garimidi, Michael Neuder, and Tim Roughgarden, explores connections between blockchain technologies and the well-studied model of Tullock contests from game theory. They show how several protocols in use in major blockchains can be modeled and understood as Tullock contests. These applications further suggest new open problems in the theory of Tullock contests.

Our final letter, from Nikhil Garg, reflects on systems that elicit and aggregate individual preferences—such as participatory budgeting and school matching—to improve public decisionmaking. He argues that participation often skews toward more advantaged individuals, and calls for the design of mechanisms that still learn from participants while ensuring fair outcomes for those less able to engage.

We have one research survey in this issue, from Parikshit Gopalan and Lunjia Hu. They explore recent advances in approximate calibration, which helps ensure probabilistic predictions in machine learning are reliable and interpretable for real-world decisions. They highlight challenges in defining well-behaved, computationally tractable measures of calibration error. They also present an indistinguishability perspective for understanding calibration error, and discuss implications for decision making and algorithm design.

The issue concludes with an annotated reading list contributed by Suho Shin and MohammadTaghi Hajiaghayi. They overview the recent, lively activity on the *delegation* problem. In this classical problem from microeconomic theory, an uninformed decisionmaker seeks to make a choice of action, and designs a mechanism to

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

delegate this choice to a more-informed agent. They overview the many interesting variants of the problem, as well as connections to well-loved models in EconCS such as prophet inequalities and Pandora's box.

As always, we would like to thank communications chair Yang Cai and technical lead Jinzhao Wu for their help publishing the issue. Please continue to volunteer letters, surveys, annotated reading lists or position papers; your contributions are what keep the Exchanges a lively venue for the SIGecom community. We hope you enjoy this issue.

Letter from the SIGecom Executive Committee

MICHAL FELDMAN (chair)
Tel Aviv University
and
FEDERICO ECHENIQUE (vice chair)
UC Berkeley
and
BRENDAN LUCIER (secretary-treasurer)
Microsoft Research New England

It's an honor for us to serve as the SIGecom Executive Committee. As the incoming leadership team, we began our term with a deep appreciation for the strong and diverse community that has grown around the intersections of economics and computation. Our goals are to support the SIG's continued growth while emphasizing interdisciplinary engagement, recognition of impact, community involvement, and encouraging diversity across a range of dimensions.

Our flagship event, the ACM Conference on Economics and Computation (EC), continues to grow and thrive. EC'24 was hosted at Yale University this past summer and marked the 25th anniversary of the conference. With 848 submissions and 204 accepted papers, EC'24 broke all previous records.[1] The conference featured outstanding work across theory, empirics, and applications. We are grateful to General Chair Dirk Bergemann, PC Chairs Bobby Kleinberg and Daniela Saban, and the many organizers and volunteers who helped make EC'24 a success. The conference was accompanied by a robust workshop program, a special session on highlights beyond EC, along with a virtual preview week that featured the annual mentoring workshop and many high-quality tutorials.

The continuing growth of EC was especially apparent in the EC'24 town hall, which was a lively event with participants from many different backgrounds. This growth is exciting but also comes with challenges. The large number of submitted papers raises questions about both the reviewing process and the length and size of the conference. We are in continuing conversations with each year's organizers to consider how to deal with these challenges. As always, we invite you to reach out to us with any thoughts—EC is your conference and we want it to be as successful and enriching as possible!

The SIGecom Winter Meetings have continued to grow into a vibrant forum for focused discussions. The 2024 meeting, co-organized by Sigal Oren and Ran Shorrer, centered on Behavioral Models and brought together economists and computer scientists to explore models of behavioral agents in economic environments. The 2025 meeting, co-organized by Renato Paes Leme and James Wright, focused on Market

[1]Editors' Note: This letter was prepared just prior to EC 2025.

Authors' addresses: mfeldman@tauex.tau.ac.il, fede@econ.berkeley.edu, brlucier@microsoft.com

Fig. 1. The EC'24 town hall meeting.

Design and Generative AI. Through a combination of invited speakers, contributed talks, and a fireside chat, it explored research ideas and cross-disciplinary topics in this exciting and developing domain.

Internally, we've put particular focus toward improving continuity and support for the many invaluable volunteers who help run SIGecom events and activities. One major effort is documenting best practices and institutional knowledge for all major SIGecom roles. While still in development, these expanded "how-to" documents are already helping incoming organizers build on their predecessors' work.

This year also marks the formal establishment of a SIGecom Communications Committee, tasked with supporting our online presence and community outreach. The committee is chaired by Yang Cai, with Kira Goldner serving as Social Media Lead and Jinzhao Wu as Technical Lead. We are excited about this team's energy and ideas, and look forward to expanded communications across multiple platforms. We also want to extend our sincere thanks to Yannai Gonczarowski, who served as Information Director with dedication and creativity over the past several years, and helped lay the foundation for this new team.

Looking ahead, we are forming a Fundraising Committee to make the process of seeking and managing institutional sponsorships more organized and sustainable. Our goal is to ensure smoother transitions, better documentation, and stronger institutional memory in this crucial area. We hope in particular to foster long-term relationships with industrial partners that can help with the financial health of EC and the rest of our activities.

We'd like to take this opportunity to reinforce our ongoing call for Special Initiative proposals. SIGecom has annual discretionary funding available to support

community-building efforts such as mentoring programs, inclusion efforts, and other pilot events. Past initiatives have included the annual EC Mentoring Workshop, EC childcare support funding, and the SIGecom Winter Meetings. We encourage members to propose creative new initiatives that further SIGecom's mission.

We are deeply grateful to the outgoing Executive Committee—Nicole Immorlica, Scott Kominers, and Katrina Ligett—for their outstanding service over the previous term. They navigated the SIG through the challenges of the pandemic and helped strengthen our community through initiatives in diversity, communication, and hybrid engagement. Their leadership left the SIG stronger, more connected, and better prepared for the future.

We're also thankful for the many volunteers who contribute to SIGecom's success in both visible and behind-the-scenes roles. The SIG could not function without you! As always, we invite members of our community to get involved. Whether you're interested in organizing an event, submitting a nomination for one of our awards, launching a special initiative, or serving in a formal role, we welcome your participation. Thank you for the opportunity to serve.

SIGecom Winter Meeting 2025 Highlights

BAHAR BOROOMAND
University of Alberta
and
SAFWAN HOSSAIN
Harvard University
and
EDEN SAIG
Technion – Israel Institute of Technology

Bahar Boroomand is a M.Sc. student in Computing Science at the University of Alberta, passionate about Machine Learning and its applications. Her current research focuses on alleviating biases caused by rating-based scoring algorithms in recommender systems using data-driven machine learning techniques.

Safwan Hossain is a PhD Candidate in Computer Science at Harvard University. His research interests are broadly at the intersection of economics and computer science, involving questions related to strategic behavior, fairness, and incentives that arise in supervised or online learning settings with multiple agents. Prior to his PhD, Safwan received his BASc. in Electrical Engineering and MSc. in Computer Science from the University of Toronto, and spent two years working as a machine learning engineer at Cerebras Systems.

Eden Saig is a PhD candidate in Computer Science at the Technion, advised by Nir Rosenfeld. His research focuses on machine learning and algorithmic decision-making in social contexts, aiming to develop socially favorable learning algorithms for behavioral environments with dynamics and incentives. Before starting his PhD, Eden received a BSc in Computer Science, BSc in Physics, and an MSc in Computer Science, all from the Technion.

General Terms:

Additional Key Words and Phrases:

---

1. INTRODUCTION TALKS

1.1 Haifeng Xu: Rethinking Online Content Ecosystems through the Lens of Computational Economics

The first invited talk of the session, by Haifeng Xu from the University of Chicago, highlighted a new research agenda: studying the wide range of problems in online content ecosystems through the formalisms of computational economics. Online content recommendation engines—core to platforms like YouTube, Instagram, and TikTok—serve personalized content to billions of users daily. The classic model considers both the users and the content library to be static, with the recommendation engine responsible for generating a mapping between the two. Xu's talk envisions a richer model that incorporates the incentives of content creators (e.g., YouTube rewarding videos based on length and views), the myopic and dynamic behavior of

---

consumers, and the increasingly prominent role of AI in both generating content and being trained on it. This is a rich, dynamic multi-agent environment and the remainder of the talk considers two distinct directions within this framework:

(1) Diagnosing and optimizing existing content ecosystems
(2) How AI-generated content can transform future content ecosystems

Existing content ecosystems can be seen as a two-sided market between self-interested consumers and creators, with the platforms acting as a powerful and self-interested intermediary. Recent works have studied parts of this interaction. Several recent works study the "supply-side" interactions between creators, who generate traffic, and platforms, who benefit from this traffic and share revenue. These include understanding, among others, creator competition [Ben-Porat and Tennenholtz 2018; Ben-Porat et al. 2020], incentivized matching mechanisms [Mladenov et al. 2020], and content distribution at equilibrium [Jagadeesan et al. 2023]. Less studied is the "demand-side", which model interactions between platforms and users. [Kleinberg et al. 2024] focus on improving recommendations through "behavior-aware" system learning. Modeling the overall ecosystem with all three types of players is understudied. Of note is [Yao et al. 2023]: they study mechanisms to incentivize content creation for user welfare maximization under a self-interested platform. The proposed mechanism ends up introducing more competition for congested topics. Importantly, a variant of the mechanism was tested and validated on Instagram Reels, involving over ten million users and creators.

Xu suggests that the rise of powerful AI systems constitutes a *fourth* player within this ecosystem. AI systems can act as content generators and thus compete with human creators. In turn, they also rely on platforms and user feedback to train and validate their models. Each of these roles/perspectives lead to numerous unexplored research questions and can fundamentally alter the dynamics of content ecosystems. [Taitler and Ben-Porat 2025], for instance, study the AI-creator-platform dynamic and suggests that AI systems can strategically give worse answers to allow for more high-quality human generated content in the short term in order to increase their long-term utility. They also observe a Braess paradox phenomenon occurring once AI systems partake in content generation. [Raghavan 2024] studies the AI-consumer interaction and suggests that it may lead to reduced content diversity. [Duetting et al. 2024] studies the AI-platform-consumer interaction and illustrates how AI systems can be part of new monetization mechanisms. Overall, the talk concludes by stressing that "incentives and agency are crucial to both learning algorithms and market mechanisms for resolving these pressing issues".

1.2 Jon Kleinberg: Language Generation in the Limit

In his intriguing talk, Jon Kleinberg presented a formal abstraction which aims to capture the foundational properties of generative AI [Kleinberg and Mullainathan 2024]. He began by asking whether there exists a simple theoretical metaphor — analogous to the metaphor of "Alice and Bob" in secure communication, or the metaphor of "Byzantine generals" in distributed systems — which captures the core properties of generative AI and enables rigorous analysis. Towards this, Kleinberg proposed framing the task of "learning to generate" as an algorithmic question:

Can an algorithm, presented only with a stream of valid words of some formal language, eventually start emitting never-before-seen words of that language?

To formalize this question, [Kleinberg and Mullainathan 2024] extend the classical framework of *language learning in the limit*, initially formulated by [Gold 1967], and further characterized by [Angluin 1979; 1980]. In the framwork, a language L is a countably infinite set of words, and there is a countable set of languages $\{L_1, L_2, \dots\}$. An adversary initially selects a target language L_k from that set, and interacts with an algorithm over discrete time steps. At each step, the adversary reveals a previously unseen word $w \in L_k$, the algorithm emits an output, and no further information about L_k is provided except for these positive examples.

In the original Gold-Angluin framework, the output of the algorithm at each step is a guess about the index of the target language L_k, and the goal is design an algorithm which stops making mistakes after a finite number of steps. The classic result of [Gold 1967] shows that this task is impossible in general, as an adversary could construct word streams for which the algorithm makes an infinite number of mistakes. However, when shifting focus from language identification to language generation, [Kleinberg and Mullainathan 2024] reveal a fundamental contrast: They present an algorithm that, in the limit, produces an infinite stream of valid and previously unseen strings from the target language L_k, despite not being able to explicitly identify it in the Gold-Angluin sense.

The algorithm relies on the definition of *language criticality*, which identifies progressively thinner languages consistent with the data seen so far. At each step, the generation algorithm maintains the critical language, and generates a previously-unseen word from it. While this guarantees validity in the limit, the definition of criticality also implies that each critical language is a strict subset of the previous ones. Thus, the algorithm may reach a state where the critical language is a strict subset of L_k, preventing it from generating all possible words in the target language. This reveals a trade-off between validity and breadth: to avoid mistakes in generation, the algorithm must permit incomplete coverage of the target language. Interestingly, this trade-off draws qualitative parallels to linguistic phenomena observed in practice, such as vernacular adoption dynamics in online communities [Danescu-Niculescu-Mizil et al. 2013], and quality-diversity tradeoffs in LLMs.

Beyond their main result, [Kleinberg and Mullainathan 2024] provide stronger convergence guarantees for finite sets of languages, and extend the framework to settings with prompting. Subsequent work has already begun exploring different aspects of the validity–breadth trade-off [Charikar and Pabbaraju 2024; Kalavasis et al. 2024a; 2024b; Kleinberg and Wei 2025], extending the stronger convergence guarantees to certain infinite sets of languages [Li et al. 2024], and exploring interaction models with noisy examples [Raman and Raman 2025]. Each line of inquiry provides new perspectives on the fundamental properties of language generation, and creates intriguing frontiers for future work.

1.3 Manish Raghavan: Competition and Diversity in Generative AI

In his talk, Manish Raghavan explored the tension between competition and diversity in the context of generative AI, drawing attention to a growing concern: generative models become ubiquitous across many domains, but the outputs they produce remain relatively homogeneous. This phenomenon, which relates the the

general notion of *algorithmic monoculture* [Kleinberg and Raghavan 2021], arises when many individuals rely on the same language model, leading to results which are less diversified. For instance, while AI tools may enhance individual productivity in tasks such as brainstorming or ideation, they might also increase homogeneity by guiding users toward similar answers. This motivates a natural question: Can we design environments that encourage novelty alongside correctness?

Towards this, [Raghavan 2024] introduces a stylized game-theoretic model to study this question. The model defines a game over n players, where each action is a categorical distribution over outputs with an ordering constraint, representing the output distribution of an LLM given some prompt. When multiple players have the same realized output, they split the reward, reflecting competition for audience attention or market share. The theoretical analysis shows that stronger competition induces players to adopt more diverse strategies, although equilibrium behavior remains less diverse than the social optimum. Perhaps surprisingly, the relative ranking of different strategies depends on competitive intensity, and a generative model that has the best performance in isolation can become suboptimal in the presence of competition due to lack of diversity.

Empirical validation is performed through simulations of the game Scattergories, played by LLMs under two settings: one where players share the same language model but choose generation temperatures strategically, and another where they can also choose which model to use. The results demonstrate that the best sampling strategy depends not only on the temperature but also on the specific model and the number of players. Models better at sampling from the tails of their output distributions had greater diversity in their outputs, and performed better as competitive pressure increased.

The talk concluded with several takeaways and open questions. While generative AI tools hold immense promise, their widespread adoption risks diminishing diversity. Competition, both between users and between models, can act as a force to counteract this. This points to a broader design question for AI systems: We often optimize systems for correctness, but can we optimize for novelty and diversity? As AI-generated content permeates more aspects of society, understanding and shaping these dynamics will be a vital challenge for both theorists and practitioners.

1.4 Yannai Gonczarowski: Algorithmic Collusion by Large Language Models

Yannai Gonczarowski presented a talk on his recent paper of the same title, co-authored with Sara Fish and Ran Shorrer [Fish et al. 2024]. He begins by defining the classical notion of collusion in economics: traders/competitors meeting to jointly raise the price of a certain good, at the expense of the public. He comments that in an increasing number of settings, automated AI driven agents are being used for pricing, and this work formally explores the potential for autonomous algorithmic collusion when large language models (LLMs) are used for this task. Three main questions are addressed within this context:

(1) Are LLMs good at pricing tasks?
(2) If multiple firms separately use LLMs for pricing, can this lead to supra-competitive prices?
(3) What mechanisms promote or prevent collusion?

Using a repeated Bertrand oligopoly environment, the authors first demonstrate that LLMs are capable of learning near-optimal monopoly pricing quickly and reliably in a monopoly setting. The pricing agent here does not require any fine-tuning and is instead based completely on in-context information. This includes the task prompt, basic information about the instance, the market history (past prices, quantities sold, profits earned etc), and past reasoning stated by the LLM. GPT-4 converges to the optimal monopoly price in all settings, with other LLMs showing more varied performances.

The duopoly setting is considered next with the precise question: if two firms are using GPT-4 for pricing, does it lead to competitive or supra-competitive pricing? Naturally, the in-context prompt used for pricing matters deeply, and the work considers two variants: (1) explicitly mentioning the LLM to not undermine profitability, and (2) mentioning that lower pricing than competitors will lead to higher sales volume. While both prompts lead to supra-competitive pricing, prompt (1) leads to higher prices than prompt (2). To understand the pricing process further, the LLM reasoning is analyzed, and evidence illustrates that the agents are concerned about avoiding a price war, especially under prompt (1).

The talk underscores the regulatory challenges posed by LLMs: their ability to autonomously adopt collusive strategies even under benign instructions, their black-box reasoning, and the sensitivity of outcomes to prompt wording. Unlike traditional Q-learning agents, the basis of past automated pricing works, LLMs are pre-trained, adaptable, and readily deployable, exacerbating concerns over the real-world applicability of algorithmic collusion. Reevaluating regulatory frameworks in light of these findings is fundamental.

1.5 Sanmi Koyejo: On Shaping AI's Impact on Billions of Lives

In his talk, Sanmi Koyejo presents a multifaceted vision for developing artificial intelligence technologies that maximize societal benefit while mitigating harm. He argues for reorienting AI development toward the public good by embedding economic, ethical, and sociotechnical considerations into the design and deployment of AI systems. The underlying premise is that the default trajectory of AI, driven largely by market incentives, may not align with broader societal interests unless interventions are made deliberately and early.

A core theme of the talk is the human-AI collaboration paradigm. Rather than envisioning AI as a replacement for human labour, Koyejo advocates for building synergistic systems that augment human capabilities, improve job satisfaction, and unlock elastic economic potential. For example, in domains like consulting, legal services, and writing, early empirical evidence suggests that AI disproportionately benefits lower-skilled professionals by narrowing performance gaps. Importantly, the speaker draws on economic theory to emphasize that AI's impact on employment will vary by sector, depending on how demand responds to increased efficiency. In areas where greater efficiency leads to increased usage, such as healthcare or education, AI has the potential to create more jobs by expanding services. In contrast, in sectors where demand remains relatively fixed, such as agriculture, efficiency gains are more likely to reduce the number of workers needed.

The talk highlights several concrete application domains where AI can be transformative. In healthcare, he discusses the potential for AI to alleviate administra-

tive drudgery and reduce burnout among clinicians, enabling them to focus more on patient care. In education, AI can help close systemic learning gaps through personalized instruction and empirically driven interventions. Koyejo stresses the importance of continuous evaluation and measurement infrastructure in both domains, drawing parallels to high-stakes fields like clinical trials. A key insight is the need for a shift from static, pre-deployment evaluation to dynamic, post-deployment monitoring, a critical requirement given the evolving nature of AI systems and their societal impacts.

He also devotes considerable attention to the information ecosystem, where trust, polarization, and misinformation present pressing challenges. He identifies the dangers of overtrust in AI-generated content, especially in the context of natural language interfaces, and proposes mechanisms for calibrated trust, such as user-facing confidence indicators, citation linking, and interpretable model diagnostics.

Rather than prescribing a single moonshot, Koyejo calls for a portfolio of milestone-driven efforts, ranging from targeted prize challenges to the establishment of interdisciplinary research centers. He encourages researchers to contribute to foundational infrastructure. This pluralistic approach reflects the belief that shaping AI's impact on billions requires collective, iterative innovation rather than top-down mandates. Policymakers, technologists, and civil society actors are urged to co-create governance mechanisms that are legally grounded yet adaptable to the unique demands of AI.

2. FIRESIDE CHAT WITH PRESTON MCAFEE AND PRABHAKAR RAGHAVAN

The 2025 Winter Meeting featured a thought-provoking Fireside Chat between Preston McAfee, Google Distinguished Scientist and a pioneering expert in auctions, market design, and computational economics, and Prabhakar Raghavan, Google's Chief Technologist and a renowned authority on search, algorithms, and web-scale systems. Drawing on decades of influential research and leadership across academia and industry, McAfee and Raghavan engaged in a dynamic conversation about how AI could influence market behavior, support proof automation in microeconomic theory, expose limitations in current macroeconomic modelling, and introduce new approaches to reasoning about complex socio-economic systems.

Preston McAfee, Google Distinguished Scientist, is an expert on pricing, auctions, antitrust, business strategy, market design, computational advertising, and machine learning applied to exchanges. He has published over 130 refereed articles, holds eleven patents, and has authored three books. His research notably influenced spectrum auction design, earning him the Golden Goose award. After earning his B.A. from the University of Florida and his Ph.D. in economics from Purdue University, McAfee spent 28 years as a professor at UWO, UT Austin, and Caltech. He held leadership roles at Yahoo!, Google, and Microsoft, including Chief Economist at Microsoft. In 2006, he published the open-access textbook Introduction to Economic Analysis, awarded the SPARC Innovator Award in 2009.

Prabhakar Raghavan is Google's Chief Technologist and one of the foremost authorities on algorithms and web search. He is the co-author of the foundational texts Randomized Algorithms and Introduction to Information Retrieval. Prab-

hakar has published over 100 papers and holds 20 patents, particularly in link analysis. At Google, he served as Senior VP for Knowledge & Information, overseeing products like Search, Ads, and Gemini. Before Google, he led Yahoo! Labs, served as CTO of Verity, and spent 14 years at IBM Research. He holds a Ph.D. from UC Berkeley and a B.Tech from IIT Madras. Prabhakar is a member of the National Academy of Engineering, a Fellow of ACM and IEEE, and recipient of the 2017 WWW Test of Time Award.

How do you think AI will influence markets beyond traditional concerns like collusion?

Prabhakar. One promising area lies in using reinforcement learning to augment mathematical proofs, particularly in theoretical computer science and microeconomics. Recent efforts have explored automating proof discovery for hardness of approximation results by learning to optimize the "gadgets" that underpin such proofs. While these AI systems haven't yet proven new theorems, they have independently rediscovered known ones using novel constructions not present in training data. These methods, if extended to combinatorial auctions and mechanism design, may refine classical hardness results and reduce proof complexity, but they still face challenges in validation and formal proof checking.

Preston. The macroeconomic side presents deeper methodological challenges. Traditional economic counterfactuals, such as Fogel's analysis of GDP without railroads, relied on constructing plausible substitutes to estimate upper bounds. This logic can be adapted to AI's economic impact by asking what it would cost to replicate AI's functionality via non-AI means, but it remains flawed, since the bundle of activities changes when costs drop. AI shifts the equilibrium by enabling behaviors that weren't previously feasible. The difficulty lies in modelling these dynamic substitutions and interdependencies across sectors.

Can LLMs or RL-based systems meaningfully contribute to macroeconomic modeling?

Prabhakar. While LLMs encapsulate vast textual knowledge, they reflect how people write about behavior rather than how they act. This makes them imperfect for modelling human strategy. However, at the aggregate level, macro behavior is often smoothed out, allowing some usefulness in high-level prediction. Drawing inspiration from DeepMind's trajectory, from playing Atari to solving protein folding, Prabhakar suggested that macroeconomic simulations could eventually be framed as multi-agent reinforcement learning environments. Agents could evolve over repeated rounds, discovering stable strategies akin to economic equilibria.

Preston. Indeed, modelling economies requires hybrid systems, treating some actors as markets and others, like corporations or key individuals, as decision-makers. While firms understand their supply chains, they struggle to model systemic interdependencies. Here, AI could help by simulating how individual decisions propagate through complex global trade networks. With trade integration now twice as deep as in 1928, understanding these chains is essential, especially amid rising protectionism. AI may offer the only scalable way to capture such emergent, nonlinear effects.

How might qualitative or textual data enhance traditional economic models?

Prabhakar. Language models can act as transformation systems: converting internal events or raw data into polished narratives, and potentially reversing that process. This opens up opportunities to extract soft signals, like executive churn or tone of announcements, and feed them into economic forecasts. For example, capturing why certain cars under- or over-perform in sales, despite technical specifications, may hinge on media coverage and public perception. These hybrid models blending structured and unstructured data could redefine how economists model demand or investor response.

Preston. Traditional models underweight soft signals because they are hard to quantify. Events like major leadership changes currently impact stock prices through gut reactions. But LLMs offer a path to formalizing these signals. Mapping unstructured news into structured risk assessments or demand adjustments could allow for richer, more sensitive models. This is especially valuable in markets where sentiment and narrative matter as much as measurable fundamentals.

What are the broader implications for modelling and equilibrium analysis in AI-influenced systems?

Preston. In practice, economic behavior often deviates from equilibrium. At Yahoo, advertising markets rarely settled into static outcomes. Instead, they evolved through reactive strategies. This mirrors the behavior of generative adversarial networks, which approximate equilibria not through optimization but via iterative best responses. Evolutionary dynamics, rather than rational-agent assumptions, may offer more realistic economic models, albeit harder to construct. These models could better fit real data and support policy decisions in complex, adaptive systems.

Prabhakar. Looking ahead, the key question is whether AI-infused macroeconomic analysis will influence actual policymaking. It's one thing to critique policy papers using LLMs or propose speculative models; it's another to change how governments approach economic strategy. The hope is that in the next 10–15 years, these tools won't just enrich analysis but reshape how economic decisions are made, grounding policy in richer, AI-assisted modelling that bridges qualitative and quantitative domains.

REFERENCES

Angluin, D. 1979. Finding patterns common to a set of strings. In *Proceedings of the eleventh annual ACM Symposium on Theory of Computing*. 130–141.

Angluin, D. 1980. Inductive inference of formal languages from positive data. *Information and control 45,* 2, 117–135.

Ben-Porat, O., Rosenberg, I., and Tennenholtz, M. 2020. Content provider dynamics and coordination in recommendation ecosystems. *Advances in Neural Information Processing Systems 33*, 18931–18941.

Ben-Porat, O. and Tennenholtz, M. 2018. A game-theoretic approach to recommendation systems with strategic content providers. *Advances in Neural Information Processing Systems 31*.

Charikar, M. and Pabbaraju, C. 2024. Exploring facets of language generation in the limit. *arXiv preprint arXiv:2411.15364*.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. 307–318.

Duetting, P., Mirrokni, V., Paes Leme, R., Xu, H., and Zuo, S. 2024. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*. 144–155.

Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. 2024. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806 7*.

Gold, E. M. 1967. Language identification in the limit. *Information and control 10,* 5, 447–474.

Jagadeesan, M., Garg, N., and Steinhardt, J. 2023. Supply-side equilibria in recommender systems. *Advances in Neural Information Processing Systems 36*, 14597–14608.

Kalavasis, A., Mehrotra, A., and Velegkas, G. 2024a. Characterizations of language generation with breadth. *arXiv preprint arXiv:2412.18530*.

Kalavasis, A., Mehrotra, A., and Velegkas, G. 2024b. On the limits of language generation: Trade-offs between hallucination and mode collapse. *arXiv preprint arXiv:2411.09642*.

Kleinberg, J. and Mullainathan, S. 2024. Language generation in the limit. *Advances in Neural Information Processing Systems 37*, 66058–66079.

Kleinberg, J., Mullainathan, S., and Raghavan, M. 2024. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Management science 70,* 9, 6336–6355.

Kleinberg, J. and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences 118,* 22, e2018340118.

Kleinberg, J. and Wei, F. 2025. Density measures for language generation. *arXiv preprint arXiv:2504.14370*.

Li, J., Raman, V., and Tewari, A. 2024. Generation through the lens of learning theory. *arXiv preprint arXiv:2410.13714*.

Mladenov, M., Creager, E., Ben-Porat, O., Swersky, K., Zemel, R., and Boutilier, C. 2020. Optimizing long-term social welfare in recommender systems: A constrained matching approach. In *International Conference on Machine Learning*. PMLR, 6987–6998.

Raghavan, M. 2024. Competition and diversity in generative ai. *arXiv preprint arXiv:2412.08610*.

Raman, A. and Raman, V. 2025. Generation from noisy examples. *arXiv preprint arXiv:2501.04179*.

Taitler, B. and Ben-Porat, O. 2025. Selective response strategies for genai. *arXiv preprint arXiv:2502.00729*.

Yao, F., Li, C., Sankararaman, K. A., Liao, Y., Zhu, Y., Wang, Q., Wang, H., and Xu, H. 2023. Rethinking incentives in recommender systems: are monotone rewards always beneficial? *Advances in Neural Information Processing Systems 36*, 74582–74601.

EconCS in Industry: Skills to Succeed as an Applied Scientist

NIKHIL R. DEVANUR
Meta
and
RENATO PAES LEME
Google
and
OKKE SCHRIJVERS
Central Applied Science, Meta

In our years as applied scientists and managers at Google, Amazon, and Meta, we have seen both the strengths that EconCS researchers can leverage in industry, as well as common challenges that these researchers face. Most EconCS PhD programs do not emphasize exploratory data analysis, applied machine learning and statistics, or a coding mindset, even though these are valuable skills to have in industry. In this article we share how these skills are leveraged, and how you can invest in building these skills now. In doing so, we hope to make it easier for people from the EconCS community to be successful in industry, be it during an internship, a sabbatical, as a part-time consultant, or as a full time applied scientist!

1. INTRODUCTION

EconCS is Multidisciplinary

While the EconCS field has its roots in Theoretical Computer Science, since its very inception, the field has spanned academic disciplines. When Nisan and Ronen published "Algorithmic Mechanism Design" in 1999, they connected the fields of economics and computer science together by focusing on possible strategic input to algorithms. Over the last 25 years the field has also seen significant cross-pollination between academia and industry: ad auctions, autobidders, blockchains and now foundation models have spurred significant theoretic work, and conversely these new theoretical results have shaped the product offerings that tech companies have developed. Applied scientists at tech companies operate at the intersection of this and facilitate a two-way street between academia and industry. They have a deep understanding of the product problems that exist and help formalize and popularize such problems in academia, and they leverage the latest developments in academia to drive impact on the companies' products.

Industry is rewarding

There is a large overlap between the types of problems academics and applied scientists work on, but there are important differences too. Applied scientists spend a good amount of time working directly with product teams to understand the in-

Authors' addresses: ndevanur@meta.com, renatoppl@google.com, okke@meta.com

tricacies of the problem space, which helps in formulating theoretical models that capture the most important factors while abstracting away from peripheral concerns. When subsequently developing algorithms or methods for these models, it's key to test this out in the actual production system to confirm that the assumptions and abstractions were indeed justified, and to make sure that the new change actually has a meaningful impact on users. Eventually, the algorithms they work on may impact millions or even billions of end users!

EconCS skills do well in industry

EconCS training provides applied scientists a unique perspective of seeing production systems from the perspective of incentives that act on participants. One classic example is ad auctions, where every modification to the auction changes the behavior of both advertisers bidding in such a system as well as the users seeing those ads. Advertisers may lower bids when presented with higher prices and users may click less when presented with lower quality ads. The same applies in many other settings: a change in the content ranking algorithm of a user's feed, leads to creators producing different content. And given an algorithmic change that decides who gets priority for scheduling jobs in a shared computing platform, users will find ways to improve their own chances of getting allocated by gaming the system.

It is often the case that a new algorithm performs very well when we simulate it under current user behavior but when we deploy, the user behavior changes and we end up in a worse place than we were originally. The EconCS perspective of reasoning about equilibrium and incentives can help identify ways in which real world systems can be gamed and mechanism design, information design and social choice can provide methods for making such systems more robust to manipulations.

What's the problem?

While applied science in industry is exciting and EconCS researchers have advantages in utilizing their skillset in industry, the three authors have found that there are also challenges to making the transition, and in our years have seen those challenges come up for others as well. Since most EconCS researchers have a background in theoretical computer science, many are less familiar working with real data, which makes it significantly more difficult to develop the product understanding that's necessary to develop good theoretical models and algorithms for them. Additionally, almost all roles in tech companies require good coding and ML engineering skills. Even when working with other software engineers (SWEs) and machine learning engineers (MLEs), there's a real benefit in being able to develop prototypes and initial ML models yourself to prove out the methods that you're proposing. It's possible to be a highly successful EconCS researcher in academia without picking up these skills, but learning these skills are key to making a smoother transition to applied science in a tech company.

How will this article help solve the problem?

In this article we discuss the areas that we have found to be most beneficial that a typical EconCS PhD program may not include in their curriculum: exploratory data analysis, machine learning and statistics, and a coding mindset. For each of these areas, we share how these areas show up in the work, what recommended

resources are to build the skills, and actionable advice on how you can build the skills. In writing this article, we hope to give more insight into what working in industry is like, and give people all the resources to prepare for an industry research career. Industry research can be a really rewarding career; hopefully this article will help smooth the transition for those who are excited to pursue it!

2. EXPLORATORY DATA ANALYSIS

The problem with real-world problems is that they tend to be messy. It's all well and good to design an auction algorithm that works well for Myerson-regular distributions (the authors of this article have all done this), but how do you know if this is a reasonable assumption when you're dealing with real-world input? A key tool in an applied scientists toolbox is the ability to grab real data and work with it. One of the most useful basic skills in this domain is Exploratory Data Analysis, or EDA.

Exploratory Data Analysis is the iterative process of learning properties of the data that you're working with. Typically you start with a question or hypothesis (such as: the bids in an auction come from a Myerson-regular distribution), you then summarize, visualize or model your data, and finally you use what you learn to ask new questions about the data. For example, you may find that the aggregate distribution of bids is bimodal (which isn't Myerson-regular), prompting the question if there are two different populations in the dataset, each corresponding to one of the peaks. In this process you'll likely find that your data needs cleaning or transformations. For example, maybe there was a production error that caused bids to be logged as NULL.

While it may be tempting to defer to data scientists to conduct EDA, they may lack the domain knowledge to know the right questions to ask, and what are the most important take-aways from a visualization. Additionally, every additional dependency means that you are waiting for someone else's work queue to clear, leading to slower execution. By versing yourself in EDA, you'll be able to move fast, but what should you learn, and how do you get started?

Current Tools

First, it's useful to be aware of the tools that are used in industry. While these tools change over time (that's a caveat that applies to most of the things we share here in this article), they represent a good place to build the fundamental understanding and skills.

Data in industry is commonly stored in databases, and most companies will use some variant of SQL to access the data. While there are some that can work magic in SQL, generally speaking it's sufficient to know basic commands, as most of the visualization, analysis and transformations are easier done after pulling the data.

To analyze the data, there are two languages that are generally used: Python and R. While Python is generally more common, people with a background in statistics may be more familiar with R. If you're only learning one, Python is the way to go.

The Python data science toolkit is spread out over different packages: Pandas [Wes McKinney 2010] is used to represent the data and perform operations on it. It's built on NumPy [Harris et al. 2020] and in some cases, familiarity with NumPy can be helpful in data transformations. To visualize data, seaborn [Waskom 2021] is the easiest way to get quick results with minimal boilerplate code, but matplotlib [Hunter 2007] can be used for more freedom. It's typically useful to be in an interactive environment when doing EDA, for example by using Jupyter notebooks (previously called iPython). To learn more about these tools, and data analysis in Python more generally, check out the free books Python for Data Analysis [McKinney 2022] and the Python Data Science Handbook [VanderPlas 2016].

R is typically preferred by statisticians, and if you are only learning one language for data science, it should probably be Python (since it is easier to combine with general-purpose or ML code and it's more common in industry). However, because it is a more domain-specific language, R has some benefits (such as a streamlined syntax for common data processing steps). The best way to use R is to use the RStudio environment, along with the tidyverse [Wickham et al. 2019] packages. The tidyverse packages are based on the philosophy of having Tidy Data [Wickham 2014], and make it particularly easy to get the data in that format. Even if you don't plan on using R, reading the Tidy Data paper will be useful! If you want to learn R, one of the best resources is R for Data Science [Wickham et al. 2017].

Putting it to Practice

The books we referenced above have many exercises and sample datasets, and it's useful to go through them as you're making your way through the book. However, in our experience, the best way to fully develop these skills is by 1) using them to solve problems that you care about, and 2) using them consistently as part of your larger work. Different strategies may work for different people, but some suggestions that may help in achieving this are: doing an industry internship, participating in data science competitions, and including empirical sections in (some of) your papers.

A common recommendation to build these skills (be it EDA, ML, statistics, or coding) is to do an industry internship. The benefit here is that you work on a real problem, so you're building skills as a means to an end. This makes the importance of building the skills more salient and helps differentiate the parts that you'll use all the time vs the parts that are less common. We won't repeat the same recommendation in the following sections, but note that the recommendation applies there too.

For data science competitions, Kaggle is the most common one. In the majority of competitions, EDA is only the first step of the process, with a more heavy emphasis on building ML models on the data afterwards, but this can still be a great way to get hands-on experience (plus: experience with ML is also quite useful)! In addition to competitions, the site also hosts datasets (along with community code and discussions) and models.

Many theory papers don't have empirical sections or just a rudimentary one that doesn't add much insight beyond the theoretic results (we are definitely guilty of this). But that can be a missed opportunity! An empirical section can demonstrate that an algorithm can perform much better than worst-case bounds, or provide strong evidence that the conditions under which theorems are proved are reasonable. While it can be difficult to get access to data outside of interning/working in industry (and even then it can be hard to publish those datasets), there are a number of publicly available datasets that can be useful for this: Criteo has a number of datasets for online advertising, there's also a NeurIPS competition dataset [Su et al. 2024] for autobidding settings, and the Movielens dataset [Harper and Konstan 2015] can be used for general valuations. This list is not exhaustive (and biased towards our own experience); see if you can find datasets that are appropriate for your papers!

3. MACHINE LEARNING AND STATISTICS

While exploratory data analysis is all about interactively learning from data with a human in the loop, machine learning and statistics are used extensively across tech to develop algorithms automatically from past data and make decisions about which new features to launch.

Machine Learning

Machine learning is so crucial to all the tech companies that you will definitely need to work with machine learning models as part of the overall system. For example, in ad auctions a very important input into the auction is the probability of a click. We typically assume that we know the true probability, but in practice, this is the output of a machine learning model. If the model is not perfect—which it never is—that may affect the outcomes of your design. You need to be aware of the limitations of machine learning, and be aware of concepts such as overfitting and calibration.

Another common paradigm in EconCS is dealing with uncertainty, such as in the study of prophet inequalities. For such problems we either assume that we already know the distributions exactly, or that we have i.i.d. samples from an unknown distribution. In the latter case, we assume that each instance is independent, and we learn only from samples for that instance. In practice, often there are many parallel instances of the same problem and the data for all the instances are correlated. For example, in ad auctions, you can consider the auction for each keyword as an independent instance, but advertiser values for similar keywords are correlated. For instance, advertiser values for a keyword "green sweater" could go up because it is getting colder and all "sweater" related keywords are trending up, or because it is nearing St. Patrick's day and keywords for all green colored apparel are trending up. By using samples from all the keywords together, an ML model can learn such patterns.

You may need to prototype some simple ML models. For this, you need to know how to train a model using standard libraries. You need to know what features to collect, what model architecture to use, what is the loss function, and what are

typical sanity checks to run, such as normalizing the inputs. To build these skills, there are several good online courses and Python packages that may be helpful.

For many years, Andrew Ng's Coursera course has been recommended as a great starting point to learn machine learning, and for good reason! The course gives hands-on experience building ML models in Python using Numpy and scikit-learn [Pedregosa et al. 2011]. The latter (also known as sklearn) is a common ML package used for smaller scale ML training. While you won't be building production-scale models with sklearn, it's an excellent way to build sufficient familiarity with building ML models. For people that prefer learning through reading, Andrew Ng's CS229 lecture notes are also excellent. For those seeking the thrill of a competition to motivate themselves to build better models, Kaggle has a whole range of ML competitions that you can participate in.

A word on LLMs

Large Language Model technology is very exciting and rapidly evolving. For those wondering if they need to know all the ins and outs of LLM training if they're in industry, rest assured that this is not the case. However, it is useful to know how to use GenAI products, particularly for coding, which is one of the biggest use cases of these models within tech companies. In addition, we recommend thinking of creative applications of large language models in new and different areas. These tools are, for now at least, not a replacement for building coding, EDA, or ML skills, as you'll need to not only write code, but also vouch for it's correctness.

Statistics

Many important insights in practice come from understanding the behavior of the participants in your design, such as the users or the advertisers. These are typically observed and validated using large scale experiments, a.k.a. A/B tests. Often the experiments involve two sides of a marketplace, and experiment design for such marketplaces has been one of the areas where folks from EconCS have made some very interesting contributions. Most EconCS researchers are well versed in probability theory, but there are certain statistical concepts that are crucial to understand in order to analyze these experiments. These are easy to pick up, such as p-value, t-test, power analysis, winsorization, or minimum detectable effect.

There are several excellent resources to learn about how A/B tests are run in tech companies. "Trustworthy Online Controlled Experiments" [Kohavi et al. 2020] provides a recent overview of the main considerations that go into A/B testing at tech companies, written by authors who have developed these systems at Google, LinkedIn and Microsoft. "The Econometrics of Randomized Experiments" [Athey and Imbens 2017] is a more technical survey of the analysis of randomized experiments. There are situations where it is infeasible to conduct perfectly randomized experiments. "Mostly Harmless Econometrics" [Angrist and Pischke 2009] is a great guide for the most-used tools for such settings. There are no surveys specific to experimentation in markets, though interested readers can look up "budget-split experiments", "bipartite experiments" or "experimentation under interference".

4. CODING MINDSET

Coding is an essential part of being an applied scientist and can be seen as the ability to translate the algorithms we design to practice. When implementing an algorithm, one can't ignore the messiness of the real world – corner cases, data that is not properly cleaned-up or is missing, small probability events that may cause a server to crash, et cetera. Coding also forces us to confront performance issues in a very serious way: constants in the running time of an algorithm may not make a difference in a research paper but can be a deal-breaker in practice.

Knowledge of certain fundamental languages like C++ (for high performance code) or Python (for a broad swath of applications) is certainly important but specific technologies/libraries are constantly changing. So instead of focusing on specific technologies we suggest developing three things: a coding mindset, strong software engineering skills, and understanding distributed systems.

Coding Mindset

A coding mindset refers to going one step beyond algorithmic thinking, in actually seeing the algorithm through in action on real data and reasoning about which repetitive processes can be automated. Software engineering skills have to do with writing code that is tested, easy to read, maintainable by others and follows a consistent and predictable style. It is the difference between writing programs for yourself and writing code that will be later read and modified by hundreds of other engineers over a long period of time.

Software Engineering

While there are books and courses devoted to software engineering, we think there is no better way to learn than doing it in practice, by either doing an internship where you will write actual production code or contributing to open source projects where you will develop software alongside others. It is a great idea to sharpen one's coding skills using interview-preparation websites (e.g. LeetCode or HackerRank) or coding competitions (e.g. ICPC). Those help build familiarity with languages and speed in thinking about coding solutions – which can be very helpful when building rapid prototypes.

It is important to note that coding does not refer only to writing code. A very important skill is to be able to read code effectively. Documentation in industry is frequently missing, incomplete, or outdated, so the only source of truth about the behavior of a system is the code itself. An effective applied scientist should be able to interrogate the codebase to understand the behavior of a system. When we inspect the code more closely we soon find out that the common wisdom about how the system behaves may not be entirely accurate.

A second reason to learn how to read code effectively is to be able to do code reviews, i.e. verifying and vouching for code written by others. Applied researchers typically design systems that are then implemented by other engineers in the organization. It is important to be able to verify that there are no gaps between the

system designed and the system actually implemented.

Finally, when we practice reading code, we become better at writing code that is readable by others. Clear and readable code has useful comments, properly-named variables and follows a well-defined style–usually according to each company's guidelines. Well written code also contains plenty of unit-tests which verify its correctness and provide code-readers with examples on how each part of the code is used.

Distributed Systems

Most tech companies operate at a tremendous scale, serving millions and billions of users at subsecond latencies. This is enabled by complex distributed systems with many interlocking pieces. Any change that you would want to incorporate has to be a part of this overall system, so it is important to understand how these systems work, and what are the strengths and limitations of such systems. We recommend becoming familiar with the design of systems such as how a news feed is designed, or how an information retrieval system is designed. The other aspect of this is that the data that you will be working with is also large scale and typically cannot fit in the memory of a single computer. Data analysis requires working with distributed big data systems, so once again, it's good to be familiar with how to work with such systems. There are several textbooks on this, such as "Designing Data-Intensive Applications" [Kleppmann 2019] or even "System Design Interview" [Xu 2020] that can be used to learn more.

5. CONCLUSION

In this article we've aimed to give actionable recommendations for EconCS researchers that are interested in building skills for applied science in industry. A reader may go over this article and get the impression that they're woefully unprepared for a role in industry unless they do all of this! That's not the case. Many applied scientists, to a certain extent ourselves included, pick up these skills when first preparing for interviews, during internships, or on the job. However, part of the reason that so many only learn these skills at that stage, is because we simply didn't know which skills were important or how to practice them beforehand! In writing this article, hopefully we've shed some light on the skills that are valuable for applied scientists to have and how you can start building those skills right now, as part of your broader research agenda.

6. AUTHOR BIOS

Nikhil currently works in the Core Ads Growth organization at Meta. Prior to this, he was part of the Sponsored Products organization at Amazon, and even earlier, he co-founded and managed the Algorithms group in Microsoft Research, Redmond. Nikhil obtained his PhD from Georgia Tech and spent a year at Toyota Technological Institute at Chicago. He is interested in what he calls Automated Economics, which studies the question of how technology can be used to improve the efficiency of economic systems.

Renato is a Principal Research Scientist at Google Research New York, where he co-manages the Market Algorithms group in NYC. His group partners with teams

in various Google products (ads, search, cloud, ...) to apply ideas from market design to various products. Before Google, Renato obtained a PhD in Computer Science from Cornell University and did a post-doc at Microsoft Research Silicon Valley.

Okke is a Research Scientist Manager for the Experimentation and Market Algorithms team on Central Applied Science at Meta. He obtained his PhD in Computer Science from Stanford in 2017 under supervision of Tim Roughgarden and has worked at Facebook/Meta as an individual contributor and manager since then.

REFERENCES

Angrist, J. D. and Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Athey, S. and Imbens, G. W. 2017. The econometrics of randomized experiments. In *Handbook of economic field experiments*. Vol. 1. Elsevier, 73–140.

Harper, F. M. and Konstan, J. A. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst. 5,* 4 (Dec.).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. 2020. Array programming with NumPy. *Nature 585,* 7825 (Sept.), 357–362.

Hunter, J. D. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering 9,* 3, 90–95.

Kleppmann, M. 2019. *Designing data-intensive applications*. English.

Kohavi, R., Tang, D., and Xu, Y. 2020. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.

McKinney, W. 2022. *Python for data analysis: Data wrangling with pandas, numpy, and jupyter*. O'Reilly Media, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Su, K., Huo, Y., Zhang, Z., Dou, S., Yu, C., Xu, J., Lu, Z., and Zheng, B. 2024. Auctionnet: A novel benchmark for decision-making in large-scale games. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

VanderPlas, J. 2016. *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc.

Waskom, M. L. 2021. seaborn: statistical data visualization. *Journal of Open Source Software 6,* 60, 3021.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds. 56 – 61.

Wickham, H. 2014. Tidy data. *Journal of statistical software 59*, 1–23.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. 2019. Welcome to the tidyverse. *Journal of Open Source Software 4,* 43, 1686.

Wickham, H., Grolemund, G., et al. 2017. *R for data science*. Vol. 2. O'Reilly Media, Inc.

Xu, A. 2020. *System design interview: An insider's guide*. Independently published.

Tullock Contests in the Wild: Applications in Blockchains

PRANAV GARIMIDI
a16z crypto
and
MICHAEL NEUDER
Ethereum Foundation
and
TIM ROUGHGARDEN
Columbia University, a16z crypto

This letter shows how Tullock contests—a class of all-pay auctions with proportional allocation rules—can be used to model and reason about several blockchain settings. We review the fundamentals of Tullock contests and their connections to potential games. We discuss why certain properties of Tullock contests, such as sybil-proofness and compatibility with "decentralization," have made them common in blockchain applications. We illustrate how Tullock contests naturally arise in proof-of-work and proof-of-stake blockchain protocols, and are an attractive design for emerging marketplaces for blockspace and succinct proofs.

Categories and Subject Descriptors: K.4.4 [**Computers and Society**]: Electronic Commerce; C.2.4 [**Computer-Communication Networks**]: Distributed Systems

General Terms: Design, Economics, Documentation

Additional Key Words and Phrases: Tullock contests, Blockchain, Consensus mechanisms, Proof-of-Work, Proof-of-Stake, Prover markets

1. INTRODUCTION

Tullock contests have long been a valuable model for studying economic scenarios. Tullock explored the concept in [Tullock 1975; 1980] to examine the outcome of political elections. Since then, the model has found applications ranging from analyzing R&D investments in patent races [Baye and Hoppe 2003] to understanding how much sports teams spend on players [Dietl et al. 2008]. The key feature of a Tullock contest is that competitors must invest costly effort before the winner is chosen probabilistically, with winning probabilities proportional to investments. Crucially, these contests have an all-pay nature—the losing parties pay for their investments without any gain—and so participants must hedge against losing when choosing how much to invest. As a result, multiple parties invest at equilibrium and the most efficient party will not always win the context. In many blockchain-related allocation problems, there is: (i) an explicit goal of multiple active participants; and/or (ii) the possibility of sybil attacks (with one participant masquerading as many). Tullock contests are an attractive design in such settings.

Authors' addresses:
pgarimidi@a16z.com, michael.neuder@ethereum.org, tim.roughgarden@gmail.com

In this article, we first recap the basics of Tullock contests and their properties, and then proceed to a few key examples of their applications in blockchain-related settings. We highlight applications in which market designers prefer Tullock contests over more efficient alternatives due to their non-winner-take-all equilibria.

2. TULLOCK CONTESTS

A Tullock contest is an all-pay auction in which every participant (or "player") makes a costly investment, but only one wins a prize. Formally:

—There are n players.

—Each player $i \in \{1, 2, ..., n\}$ chooses an investment level $b_i \geq 0$ and pays b_i.

—The probability of winning the prize is proportional to one's investment. Thus, for investments $\mathbf{b} = (b_1, b_2, \ldots, b_n)$, player i's probability of winning (a.k.a. their *allocation*) is

$$x_i(\mathbf{b}) = \frac{b_i}{\sum_{j=1}^{n} b_j}. \tag{1}$$

In our analysis of Tullock contests, we will always assume that players have quasi-linear payoffs $U_i(x_i(\mathbf{b})) - b_i$, where U_i is an increasing, differentiable, and (weakly) concave utility function that satisfies $U_i(0) = 0$.[1]

One can interpret the outcome of a Tullock contest as allocating one unit of a divisible good to the participants at a common per-unit price (namely, the price $\sum_{i=1}^{n} b_i$). In particular, a Tullock contest is *sybil-proof*, meaning that no participant can benefit from participating under multiple identities (one cannot do better than splitting one's single-identity equilibrium bid arbitrarily over one's multiple identities).

We will see how multiple different settings can be interpreted as instantiations of this model. In this article, we consider only the complete information setting in which agents' utility functions are common knowledge, and focus on pure Nash equilibria. This is the setting studied in the relevant blockchain literature. Extending the analysis to the incomplete information setting and Bayesian-Nash equilibria, perhaps by building on the techniques in [Syrgkanis and Tardos 2013] and [Caragiannis and Voudouris 2016], is an interesting direction for future work.

2.1 Potential Games and Equilibrium Characterization

A key observation that enables the equilibrium analysis of Tullock contests is that they are *potential games*. In a potential game, the equilibrium outcomes correspond to the global maximizers of a suitably defined potential function. [Johari and Tsitsiklis 2004] give such a potential function characterizing the equilibria of Tullock contests. This function is defined on allocation vectors (rather than bid vectors), and characterizes the allocation vectors that are induced (via (1)) by equilibrium bid vectors.

THEOREM 2.1 EQUILIBRIUM CHARACTERIZATION [JOHARI AND TSITSIKLIS 2004]. *Allocations* $\mathbf{x}$ *that correspond to a pure Nash equilibrium (PNE) of a Tullock contest*

[1]We often consider the special case of linear utility functions, in which, for each i, $U_i(x_i(\mathbf{b})) = v_i \cdot x_i(\mathbf{b})$ for some value $v_i > 0$.

are exactly the solutions to

$$\max \sum_{i=1}^{n} \widetilde{U}_i(x_i)$$

subject to $\sum_{i=1}^{n} x_i = 1$ *and* $x_i \geq 0$ *for all* $i \in \{1, 2, ..., n\}$*, where*

$$\widetilde{U}_i(x_i) = (1 - x_i)U_i(x_i) + \int_0^{x_i} U_i(y)dy.$$

Proof Sketch. Each agent's best response function can be rewritten as a function of the allocation vector $\mathbf{x}$ and the total sum of bids B. Calculations then show that the first-order conditions of the optimization problem match the first-order conditions of the best-response problem faced by players. By our assumptions on players' utility functions (such as concavity), these conditions also characterize the global solutions to the optimization problem and players' best responses.

Because each utility function U_i is weakly concave, each modified utility function $\widetilde{U}_i$ is strictly concave. It follows that the potential function above has a unique maximum. While this only implies a unique equilibrium allocation, rewriting the agents' best-response bids as a function of their allocations further shows that equilibrium bids are unique. This observation gives us the following corollary.

Corollary 2.2. *In a Tullock contest, there exists a unique PNE.*

For blockchain-related applications, it will be useful to extend the basic model of Tullock costs to include player-specific cost multipliers. Consider a variation of the Tullock contest model in which player i pays $c_i \cdot b_i$ to bid b_i—in effect, some players can generate a given amount of investment more efficiently than others. This setting is equivalent to the classic Tullock contest setup in which each player i has the utility function $U_i(x)/c_i$.

Lemma 2.3. *Let* $\mathcal{T}(\mathbf{U}, \mathbf{c})$ *be a Tullock contest in which players have utility functions* $\mathbf{U}$ *and bidding costs* $\mathbf{c}$*. Then* $\widetilde{\mathbf{b}}$ *is an equilibrium bid vector for* $\mathcal{T}(\mathbf{U}, \mathbf{c})$ *if and only if* $\widetilde{\mathbf{b}}$ *is an equilibrium bid vector for* $\mathcal{T}(\mathbf{U}/\mathbf{c}, \mathbf{e})$ *where* $\mathbf{e}$ *denotes the all-ones vector.*

We omit the straightforward proof.

2.2 Equilibrium analysis

Tullock contests have two defining characteristics: (i) the all-pay nature in which players pay their bid even when they lose, and (ii) the proportional allocation. Combined, these characteristics imply that agents have strictly diminishing returns on investment (even with linear utility functions). For this reason, equilibria of Tullock contests are generally oligopolistic outcomes. For example, when agents have linear utility functions, the agents with larger utility functions receive higher allocations at equilibrium, but the allocation is split over two or more agents. It follows that Tullock contests do not generally implement fully efficient equilibria. [Johari and Tsitsiklis 2004] quantify equilibrium inefficiency in Tullock contests and show that the worst-case "price of anarchy" is precisely 3/4—the sum of agent utilities at equilibrium in a Tullock contest is always at least 75% of the maximum possible, and this bound is tight in the worst case.

Theorem 2.4 Efficiency of Tullock contests [Johari and Tsitsiklis 2004]. *In a Tullock contest, let* $\mathbf{d}^*$ *be the welfare-maximizing allocation and* $\mathbf{d}$ *the allocation corresponding to the unique PNE. Then:*

$$\sum_i U_i(d_i) \geq \frac{3}{4} \sum_i U_i(d_i^*).$$

Furthermore, this bound is tight: for every $\epsilon > 0$*, there exist* n *and linear utility functions* $U_1, U_2, \ldots, U_n$ *such that*

$$\sum_i U_i(d_i) \leq \left(\frac{3}{4} + \epsilon\right) \sum_i U_i(d_i^*).$$

For an alternative proof to the one in [Johari and Tsitsiklis 2004], see Section 3 of [Roughgarden 2006].

Thus, while Tullock contests generally result in inefficient equilibria, the efficiency loss is bounded. This efficiency loss may be acceptable (or unavoidable) if other considerations, such as encouraging participation to avoid centralization, are paramount.

The exact realization of the "payments" in a Tullock contest is application-dependent. For example, in Section 3.1, we discuss proof-of-work blockchain protocols, in which the payments correspond to investments in hardware and electricity. In Section 3.3, by contrast, we discuss blockspace auctions in which payments represent direct monetary transfers to an auctioneer. The efficiency objective function in Theorem 2.4 captures the utilities of players, independent of payments.

3. BLOCKCHAIN APPLICATIONS

We now turn our attention to the relevance of Tullock contests for blockchain protocols. The following simple model of a blockchain protocol suffices for this article: an ever-growing sequence of transactions, with a "leader" periodically chosen to append a new block of transactions. The leader is typically drawn from the set of physical machines running the protocol (generally called "miners" in a proof-of-work protocol or "validators" in a proof-of-stake protocol). Two common goals for "decentralized" blockchain protocols are: (i) permissionlessness, meaning that anyone should be able to participate in the protocol as a miner or validator; and (ii) no one or small group of participants should have undue control over the blockchain's transaction sequence. In part with these goals in mind, the Bitcoin protocol (among others) chooses leaders using a "proof-of-work" mechanism, repeatedly choosing a leader with probability proportional to miners' hashrates. The Ethereum protocol (among others) uses a proof-of-stake mechanism to choose leaders, with each leader chosen from the validator set with probability proportional to the amount of cryptocurrency that they have staked (i.e., locked in the blockchain protocol). We next show that these mechanisms are equivalent to Tullock contests, allowing us to use Theorem 2.1 to characterize the relative influence of different parties in these protocols. We then examine how these same ideas have been used to inspire mechanisms for new blockchain applications to achieve similar goals of having a decentralized set of participants.

3.1 Proof-of-Work Protocols

The Bitcoin protocol is "permissionless" in the sense that any party can become a "miner." Miners compete to produce new blocks of transactions by repeatedly hashing candidate strings until they find an input with a sufficiently small output value. The first miner to publish such an input is rewarded with newly created Bitcoin (currently 3.125 BTC, worth over 300,000 USD at this time of writing) and also appends a new block of recent transactions to the running transaction sequence. The threshold for "sufficiently small" is adjusted, as a function of the amount of participating hashrate, so that a new block is produced every ten minutes on average. Miners can invest in hardware and electricity to increase their hashrate and become more competitive as block producers. However, while anyone is free to make investments, only those parties that can operate the most efficiently (e.g., with access to cheap electricity) find it profitable to stay active in the network; otherwise, the cost of running the hardware exceeds the rewards they earn. Quantifying the "decentralization" of a proof-of-work protocol can then be formalized as the following question: what is the distribution of miners' hashrates at equilibrium, as a function of miners' relative costs of operation? [Arnosti and Weinberg 2022] answers this question using the framework of Tullock contests.

Model:

—There is a block reward r. (E.g., 3.125 BTC.)

—Agent i's utility is $U_i(x_i) = rx_i$. (I.e., agents are risk-neutral.)

—Agent i's cost of operating q_i units of hardware is $c_i q_i$. (E.g., reflecting hardware depreciation and electricity costs.)

—The allocation is given by $x_i = q_i / \sum_j q_j$. (The definition of proof-of-work leader selection.)

As shown in Lemma 2.3, this model is equivalent to standard Tullock contests after a simple transformation. This model implicitly assumes that hardware is homogeneous, but agents have different acquisition and/or operating costs. Equivalently, agents could have access to differing quality hardware at the same costs. The block rewards are split pro-rata according to agents' hardware, as is standard in Tullock contests. [Arnosti and Weinberg 2022] characterize the equilibrium in this setting as follows: With respect to fixed agent costs $c_1, c_2, \ldots, c_n$, define the function

$$X(c) = \sum_i \max(1 - c_i/c, 0)$$

and let c^* be the solution to $X(c^*) = 1$. Then,

Theorem 3.1 Proof-of-Work Equilibrium [Arnosti and Weinberg 2022]. *At the unique PNE of the proof-of-work Tullock contest, miners make investments* $q_i = \frac{1}{c_i}\max(1 - c_i/c^*, 0)$*, resulting in allocations* $x_i(\mathbf{q}) = \max(1 - c_i/c^*, 0)$.

Thus, for a given a cost vector, there is a threshold cost c^* such that agents with costs above the threshold do not participate at equilibrium. The following corollary provides one interpretation of this equilibrium.

Corollary 3.2 [Arnosti and Weinberg 2022]. *If miner i participates at equilibrium ($q_i > 0$), then for all j, $x_i(\mathbf{q}) \geq 1 - \frac{c_i}{c_j}$.*

This corollary demonstrates that relatively small differences in the cost of mining can result in highly concentrated allocations at equilibrium—the "natural oligopoly" referred to in the paper's title. For example, with n large and $c_i = i/(i+1)$ for each i, one can calculate $c_7 < c^* \approx .88 < c_8$ [Arnosti and Weinberg 2022]. Because $c_1 = 1/2$ and $c_7 = 7/8$, we have $x_1 \geq 3/7$, representing a substantial amount of power for a single miner.

3.2 Proof-of-Stake Protocols

In a proof-of-stake protocol (including the Ethereum protocol and many others), validators lock up capital (a.k.a. stake), and each leader is chosen with probability proportional to stake. Similarly to the Bitcoin protocol, leaders are responsible for producing blocks that append recent transactions to the running transaction sequence and are also rewarded with newly minted cryptocurrency. In a proof-of-stake protocol, the costly investment is the opportunity cost of locking up stake (as opposed to, for example, investing in U.S. treasury bills). Thus, the analysis in Section 3.1 carries over with this new interpretation, with validators choosing how much stake to invest rather than how much hardware to operate.[2,3]

An additional complication in blockchain protocols with a mature and Turing-complete smart contract layer, including the Ethereum and Solana protocols, is that validators with different levels of sophistication can earn vastly different rewards from block production. On top of standard block rewards and transaction fees, block producers can earn substantial revenue from "maximal extractable value (MEV)." Roughly, MEV refers to rents extracted by a block producer on account of their temporary monopoly power over transaction sequencing (e.g., deciding the order in which trades are executed on a financial exchange) [Daian et al. 2020]. Because some validators know about more pending transactions than others (e.g., due to business agreements with power users) and some validators are better at assembling high-MEV blocks than others (e.g., due to more computational power or better algorithms for exploring the space of possible blocks), some validators can earn much more revenue from a given block production opportunity than others.

To capture the validator heterogeneity introduced by MEV, we consider the following model:

Model:

—There is a base reward r.

—Agent i's utility for being chosen as a block producer is $U_i(x_i) = \mu_i \cdot rx_i$, with μ_i representing the agent's acumen at extracting MEV from the current block production opportunity.

—Agent i chooses an amount π_i of stake and incurs a per-unit cost of c.

[2]This analysis does not consider any returns validators earn from the stake, itself, appreciating.
[3]In practice, the majority of stake controlled by validators has been delegated to them by other parties (and so validators primarily pay operational rather than capital costs). For simplicity, in this article we'll ignore the possibility of delegated stake.

—Allocations are given by $x_i = \pi_i / \sum_j \pi_j$. (The definition of proof-of-stake leader selection.)

For example, a validator i with $\mu_i = 1$ would collect only the base reward for block production, while a validator with $\mu_i = 2$ would collect double the base reward (presumably on account of better MEV extraction). This model is mathematically equivalent to that in the previous section (see also Lemma 2.3), but the change in notation is helpful to indicate different interpretations of this model. [Bahrani et al. 2024] analyze pure Nash equilibria in this setting as a function of the relative sophistication of different validators at block production (i.e., of the μ_i's). They call a validator set (γ, k)-*competitive* if $\mu_{k+1} \geq \gamma \cdot \mu_1$; in words, at least k validators have a reward multiple that is at least a γ fraction of the largest multiple. In the context of block production, this means there are at least k parties capable of producing a block with value at least γ times that produced by the most sophisticated validator. (Larger values of k and γ correspond to "more competitive" validator sets.) [Bahrani et al. 2024] use this parameterization to upper bound the maximum equilibrium allocation of any individual validator.

Theorem 3.3 Proof-of-Stake Equilibrium [Bahrani et al. 2024]. *For every (γ, k)-competitive block producer set with $\gamma \in [0, 1]$ and $k \geq 1$, the unique PNE allocations $\mathbf{x}$ satisfy $x_i \leq 1 - \frac{\gamma k}{k+\gamma}$ for all i.*

For example, with 10 validators at least 90% as sophisticated as the most sophisticated validator, no individual validator will control more than 17.5% of the stake at equilibrium.

If one or a small number of validators are substantially more sophisticated than the rest, how can one avoid centralization (i.e., stake concentration at equilibrium)? Modern block production for the Ethereum protocol is based on *proposer-builder separation (PBS)*, a system in which validators can outsource block production to a specialized set of third parties called *block builders*. The goal of PBS is to preserve decentralization (with many validators participating at equilibrium) by confining centralization to the set of block builders.[4]

[Bahrani et al. 2024] extend their analysis to incorporate PBS, as follows. Under PBS, for each block production opportunity (called a "slot"), the corresponding leader runs a first-price auction in which block builders compete to construct the most valuable block and submit bids for their block to be chosen by the leader. (Thus, the item being sold in the auction is the current block production opportunity; the seller is the validator that was chosen as the current leader; and the bidders are the block builders.) This auction smooths out the differences in sophistication between validators, as every validator's value for being chosen as leader is now typically just the (validator-independent) revenue that they can collect as an auctioneer. Theorem 5.1 of [Bahrani et al. 2024] formally captures the effect of PBS in the above model by showing that, with PBS and at least l competitive builders, the ratio in the expected rewards obtained by any two validators for a given block production opportunity is $1 + O(1/\log l)$. That is, for large l, a block production

[4]Because builders do not participate directly in the blockchain protocol and its decisions, builder centralization is generally viewed as less concerning than validator centralization.

opportunity is almost equally valuable to all validators.[5]

This result implies that, with PBS and in the notation of our model of investments by proof-of-stake validators, the μ_i's of any two validators differ by a factor of $1 + O(1/\log l)$. Plugging this into the equilibrium analysis of Theorem 3.3 shows that, under our idealized version of PBS with n validators and l builders,

$$x_i = \frac{1}{n} + O\left(\frac{1}{\log l}\right)$$

for each validator i. That is, in this model, PBS does indeed guarantee decentralization in the validator set, despite heterogeneous validator sophistication.[6]

3.3 A Market for Block Production Rights

We now switch from analyzing the equilibria of currently implemented protocols to exploring how Tullock contests have been proposed for use in future mechanisms. We start by showing how Tullock contests can address some of the drawbacks of PBS. As discussed above, a key part to PBS helping reduce centralization in the validator set is the existence of a competitive block builder set. These builders specialize in constructing valuable blocks through many means such as unique trading strategies, business relationships, sophisticated block-building algorithms, and better networking infrastructure. The main downside to PBS is the set of builders may become centralized. In practice, it may be that only a small number of entities can consistently win block-building auctions and can reinvest those profits into gaining even more market share. At the time of writing, 96% of Ethereum blocks are built by just three different entities.[7] While block validation in Ethereum remains decentralized, builders have tremendous power in deciding which transactions are included in the running transaction sequence.

There have been discussions of alternative market structures to alleviate builders' market power and encourage participation by a larger set of builders. One widely-discussed idea is "execution tickets" [Drake and Neuder 2023], in which block production rights are allocated by lottery rather than a first-price auction. The idea is that a blockchain protocol would set a ticket price, with builders purchasing as many tickets as they wish. For every slot, the protocol would select one of the tickets uniformly at random, and the ticket owner would be granted exclusive block production rights for that slot. Payments are made up-front, and are not refunded even if the purchaser is never selected as a block producer. For the sake of this analysis, we assume that block production rights cannot be resold once the lottery winners are revealed.[8] [Neuder et al. 2024] describe how the non-resale setting is

[5]This result allows validators to build their own blocks as before (i.e., to ignore all the blocks submitted by builders) but assumes that the builders, as specialized parties, are at least as proficient at block-building as the validator. More precisely, each builder draws their value for a block production opportunity from a distribution that satisfies the monotone hazard rate condition and also first-order stochastically dominates the distribution of the validator's value for the block production opportunity.

[6]In practice, the stake distribution is also influenced by other factors, such as the different yields offered by validators to those who delegate stake to them.

[7]See `https://www.relayscan.io/` for real-time data on the Ethereum block-builder distribution.

[8]For a model that considers how this analysis changes when resale is permitted, see [Pai and

mathematically equivalent to a Tullock context, with agents' values for winning the lottery their values for block production rights and the ticket price set to \$1.

Model:

—Agent i has value v_i for block production rights, and (risk-neutral) utility function $U_i(x_i) = v_i x_i$.
—Agent i purchases b_i tickets at a cost of b_i.
—Allocations are given by $x_i = b_i / \sum_j b_j$.

In the context of execution tickets, agents' payments are direct transfers rather than implied capital or operating costs.

Theorem 2.1 can be invoked again here, replacing the μ_i's with v_i's. Define the function

$$F(x) = \sum_{i=1}^{n} \max\left(1 - \frac{x}{v_i}, 0\right)$$

and let v^* satisfy $F(v^*) = 1$. Then at the (unique) equilibrium bid vector $\mathbf{b}$, the corresponding allocations $\mathbf{x}$ satisfy

$$x_i = \max\left(1 - \frac{v^*}{v_i}, 0\right)$$

for every i.

Because execution tickets are effectively an implementation of a Tullock contest, multiple participants invest at equilibrium and have a non-zero probability of winning the block production rights for a slot. This contrasts with the "winner-take-all" nature of first-price auctions (as used in today's PBS), in which the participant with the highest value wins with certainty at equilibrium. More generally, the execution tickets design is applicable to any domain in which block production rights must be allocated, including "layer-two" protocols and shared sequencers.[9]

3.4 Proof Marketplaces

For our last example, we turn to the emerging application of marketplaces for proofs, and specifically for SNARKs (i.e., succinct noninteractive arguments of knowledge). The point of a SNARK is to enable anyone to quickly verify that a computation was carried out correctly (without redoing the computation). SNARKs are useful for a number blockchain-related applications. For example, one increasingly common architecture for a blockchain protocol is for transaction processing (and corresponding SNARK generation) to be carried out by a small number specialized "provers" (somewhat analogous to the role of builders in PBS), with the (decentralized) set of validators responsible only for SNARK verification. SNARKs are computationally intensive to produce, and proof marketplaces are designed to coordinate the clearing of the market for SNARK generation.

Proof marketplaces are two-sided markets, with agents who have demand for proofs on one side and provers on the other side. For simplicity, we consider here

Resnick 2024].

[9]For example, Espresso Systems has proposed an execution tickets-style design for a shared sequencer [Bünz et al. 2024].

a setting in which a single party demands a proof with multiple provers competing to supply it. One approach to this procurement problem would be to run a reverse first- or second-price auction. This approach runs the risk of a winner-take-all outcome, with only the most efficient prover ever producing proofs. Inspired by the successes of Tullock contests in blockchain-related applications discussed above, [Roy et al. 2024] propose a similar mechanism for proof marketplaces to address these centralization concerns. Below, we give a variation of their mechanism:

Model:

—The auctioneer (representing the buyer) posts a reward r for computing a proof ϕ.
—Each prover has a cost c_i for computing ϕ.
—Each prover submits a nonrefundable bid b_i, paid up front.[10]
—Prover i^* is randomly selected with probability $x_i(\mathbf{b}) = b_i / \sum_j b_j$.
—Upon computing and submitting ϕ to the auctioneer (which the auctioneer verifies to be a correct proof), the prover i^* is paid r.

Under this mechanism, agent i's utility is $U_i(x_i) = (r - c_i)x_i$ and thus their profit given a bid vector of $\mathbf{b}$ is

$$(r - c_i)x_i(\mathbf{b}) - b_i.$$

Thus, we get the classic Tullock contest setup in which each agent has a value of $r - c_i$ for winning the lottery and we can again use Theorem 2.1 to calculate the equilibrium. The auctioneer can expect multiple provers to compete provided r is larger than the two smallest c_i's. More generally, if provers have distinct costs, for each $k \in \{1, 2, \ldots, n\}$, there is a corresponding range of rewards for which exactly k provers will participate and receive non-zero allocations. Thus, if prover costs can be treated as common knowledge, a buyer can use the equilibrium characterization of Theorem 2.1 to choose a reward that incentivizes a target level of participation.

This mechanism for proof marketplaces relies on the fact that the auctioneer—perhaps implemented as a smart contract on a blockchain—can easily and programmatically verify the correctness of submitted proofs. The mechanism is applicable more generally to procurement problems in which satisfactory service provision can be easily verified and the auctioneer can credibly commit to paying out the reward upon successful procurement.

4. OPEN QUESTIONS AND FUTURE DIRECTIONS

—**Optimal fairness:** What does it mean for a mechanism to be optimally "fair" or "decentralized"? Is there a framework that micro-founds the optimal allocation of service providers subject to "sufficient decentralization"?
—**Optimality/uniqueness of Tullock contests:** Under a suitable metric of fairness or decentralization, are Tullock contests the best (or unique) mechanism that is fair/decentralized and also sybil-proof?

[10]In a permissionless context in which anyone can participate as a buyer or prover, these payments are burned rather than passed on to the buyer/auctioneer. (Otherwise, an agent might participate as both a buyer and a prover; by submitting a large fake bid, it could collect the reward r along with the payments made by the other provers.)

—**Guaranteed non-negative utility and sybil-proofness:** One drawback of Tullock contests is that the only way for a participant to guarantee itself non-negative utility (no matter what the other players do) is to bid 0. In particular, a participant that chooses its equilibrium bid may suffer negative utility if other participants do not, for whatever reason, choose their equilibrium bids. Is this property unavoidable for mechanisms that are sybil-proof and not winner-take-all?

REFERENCES

Arnosti, N. and Weinberg, S. M. 2022. Bitcoin: A natural oligopoly. *Manag. Sci. 68,* 7, 4755–4771.

Bahrani, M., Garimidi, P., and Roughgarden, T. 2024. Centralization in block-building and proposer-builder separation. In *Financial Cryptography and Data Security*. Springer, 331–349.

Baye, M. R. and Hoppe, H. C. 2003. The strategic equivalence of rent-seeking, innovation, and patent-race games. *Games and economic behavior 44,* 2, 217–226.

Bünz, B., Fisch, B., and Davidson, E. 2024. The espresso market design. `https://hackmd.io/@EspressoSystems/market-design`. Accessed: 2025-07-02.

Caragiannis, I. and Voudouris, A. A. 2016. Welfare guarantees for proportional allocations. *Theory Comput. Syst. 59,* 4, 581–599.

Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., and Juels, A. 2020. Flash boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. In *2020 IEEE symposium on security and privacy (SP)*. IEEE, 910–927.

Dietl, H. M., Franck, E., and Lang, M. 2008. Overinvestment in team sports leagues: A contest theory model. *Scottish Journal of Political Economy 55,* 3, 353–368.

Drake, J. and Neuder, M. 2023. Execution tickets. `https://ethresear.ch/t/execution-tickets/17944`. Ethereum Research Blog. Accessed: 2025-07-02.

Johari, R. and Tsitsiklis, J. N. 2004. Efficiency loss in a network resource allocation game. *Mathematics of Operations Research 29,* 3, 407–435.

Neuder, M., Garimidi, P., and Roughgarden, T. 2024. On block-space distribution mechanisms. `https://ethresear.ch/t/on-block-space-distribution-mechanisms/19764`. Accessed: 2025-04-21.

Pai, M. M. and Resnick, M. 2024. Centralization in attester-proposer separation. *CoRR abs/2408.03116*.

Roughgarden, T. 2006. Potential functions and the inefficiency of equilibria. In *Proceedings of the International Congress of Mathematicians (ICM)*. Vol. 3. 1071–1094.

Roy, U., Guibas, J., Pai, M., Kulkarni, K., and Robinson, D. 2024. Succinct network: Prove the world's software. `https://docs.succinct.xyz/whitepapers/succinct-network`. Accessed: 2025-03-05.

Syrgkanis, V. and Tardos, É. 2013. Composable and efficient mechanisms. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, D. Boneh, T. Roughgarden, and J. Feigenbaum, Eds. ACM, 211–220.

Tullock, G. 1975. On the efficient organization of trials. *Kyklos 28,* 4, 745–762.

Tullock, G. 1980. Efficient rent seeking. In *Toward a Theory of the Rent-Seeking Society*, J. M. Buchanan, R. D. Tollison, and G. Tullock, Eds. Texas A & M University Press, College Station, TX, 97–112.

Heterogeneous participation and allocation skews: when is choice "worth it"?

NIKHIL GARG
Cornell Tech

A core ethos of the EconCS community is that people have complex private preferences and information of which the central planner is unaware, but which an appropriately designed mechanism can uncover to improve collective decisionmaking. This ethos underlies the community's largest deployed success stories, from stable matching systems to participatory budgeting. I ask: is this choice and information aggregation "worth it"? In particular, I discuss how such systems induce *heterogeneous participation*: those already relatively advantaged are, empirically, more able to pay time costs and navigate administrative burdens imposed by the mechanisms. I draw on three case studies, including my own work – complex democratic mechanisms, resident crowdsourcing, and school matching. I end with lessons for practice and research, challenging the community to help reduce participation heterogeneity and design and deploy mechanisms that meet a "best of both worlds" north star: *use preferences and information from those who choose to participate, but provide a "sufficient" quality of service to those who do not.*

1. INTRODUCTION

A deserved point of pride for the EconCS community is the integration into everyday life the systems we have long studied, an integration often done in collaboration with researchers. In New York City, I recently voted in a participatory budgeting election and used ranked choice voting for a mayoral primary election; my neighbors submit preferences to stable matching processes that assign their children to 3-k (for three year olds), pre-k, kindergarten, middle school, and high school; and the city has embraced crowdsourcing: whenever we encounter problems as mundane as potholes or as serious as suspected lead in our water, we can submit a 311 report or request a testing kit. Each of these systems represents a triumph of an underlying community ethos: that the people have complex preferences and information of which the government is unaware, but which an appropriately designed mechanism can uncover to improve collective decisionmaking.

This article's purpose is to raise a simple, perhaps surprising, question: *is this choice and information aggregation "worth it"*? Just as democratic decisionmaking generally privileges those who (can) vote, these systems skew public resource allocation and decisionmaking in favor of those who (can) participate. And, as I will describe, substantial empirical evidence has established that participation in these mechanisms correlates with existing axes of privilege. Thus, we must ask whether the gain in information aggregation is worth the cost—or have we, in the guise of preference optimization, deployed ways to allocate scarce public resources to those best positioned to take advantage? As I will argue, this question is central to the legitimacy – and perceived legitimacy – of our systems.

My thesis is analogous to, and motivated by, those recently advanced in poli-

Author's address: ngarg@cornell.edu

cymaking, public interest technology, and behavioral economics. In their seminal book, "Administrative Burdens: Policymaking by Other Means," Herd and Moynihan [2019] argue that the information requirements to access rights such as voting and Medicare – often imposed in the name of safety, fraud detection, and choice – in practice cause people to not receive what they are entitled to. In their respective books, Schank and McGuinness [2021] and Pahlka [2023] argue that poor technology design – something as innocuous as long forms – contributes to this loss, even when well-intentioned. In "Scarcity: Why Having Too Little Means So Much," Mullainathan and Shafir [2013] explain how poverty begets poverty, because it inhibits long-term planning in favor of urgent needs. All then argue that system designers must design with this phenomenon – the time cost of participation – in mind. Analogously, I argue that effective outcomes in the face of heterogeneous participation must be a primary design goal for our field, if we want our information aggregation mechanisms to be "worth it." In other words, we should either deem "equal" participation as a necessary precondition to using choice to allocate scarce resources, or ensure that our mechanisms are robust despite heterogeneity.[1]

In this article, I first detail three case studies central to our community: complex voting mechanisms, resident crowdsourcing, and school matching. In each, I overview the promise and on-the-ground realities of how these systems affect collective decisionmaking. I highlight recent research, including my own, that has sought to understand and close the gap caused by heterogeneous participation. I then summarize shared patterns from the three case studies, including potential solutions and design principles. Finally, I overview practical and research directions on the use of choice to allocate scarce public resources. I challenge us to meet a "best of both worlds" north star: *use preferences and information from those who participate, but provide a "sufficient" quality of service to those who do not.* Simultaneously, we should help develop approaches to support balanced participation.

2. CASE STUDIES

2.1 Complex democratic mechanisms

"Equal" voting rights and participation is central to democracy. Of course, equal participation is difficult to achieve; in the United States, eligible voters who are young, lower-income, racial and ethnic minorities, or have less formal education are less likely to vote [Hartig et al. 2023]. These patterns are also present in two democratic innovations advanced in the community: participatory budgeting and deliberative democracy ("citizen assemblies").

In participatory budgeting, voters select which community projects to fund, from libraries in schools, to gym renovations, to park beautification. The Stanford Participatory Budgeting Platform has helped run over 150 elections, each of which may allocate millions of dollars [Gelauff and Goel 2024b]. New York City, Cambridge, Paris, Porte Alegro, Budapest, Helsinki, and many other cities globally all run participatory budgeting elections. At their best, these elections promise to increase

[1]I use the words, "equal," "heterogeneous," and "representative" informally. What exactly constitutes equal, or equal enough, depends on context and may be subjective. See Chasalow and Levy [2021] for a history and analysis of "representativeness" as a "foundational yet slippery concept."

civic engagement and ensure that project funding decisions are made by the people, instead of elected representatives or administrators. In deliberative democracy mechanisms, people ("panelists") are selected to deliberate, potentially over several days, over a prescribed set of issues; they are polled before and after regarding their beliefs and sometimes are tasked to make recommendations; at their best, such processes gather a diverse set of people to make decisions in accordance with what a "public sample would think if it had better conditions and information with which to explore and define the issues" [Fishkin 1991]. Such processes have been used in over 25 countries, including to make constitutional amendments in Mongolia [Lee 2024], and EconCS researchers are involved in both building online deliberation platforms and in selecting panelists [Fishkin et al. 2019; Flanigan et al. 2021].

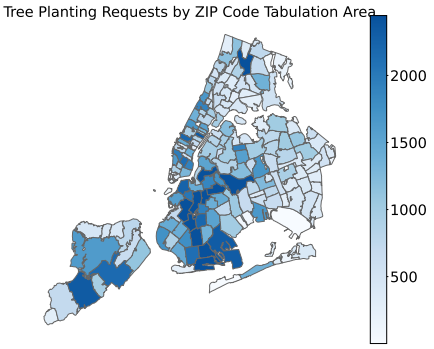
Given time costs and the use of unfamiliar methods, ensuring representative participation in these processes is a continuous, challenging task, on which researchers have rightfully focused. Participatory budgeting is often conducted online and open to all residents, including children and non-citizens; however, turnout rates are sometimes low, including at or below 5% of eligible voters, and there may be unequal participation rates by race, ethnicity, education, immigration status, and home ownership [Zepic et al. 2017; Stewart et al. 2014; Hayduk et al. 2017]. Others report that participatory budgeting increases civic engagement by otherwise disadvantaged groups [Johnson et al. 2023], and there are mixed findings on its distributive effects [Shybalkina and Bifulco 2019; Stewart et al. 2014].

Motivated by unequal participation in participatory budgeting, Gelauff et al. [2020] and Shen et al. [2021] study targeted advertising for demographically balanced participation. Gelauff and Goel [2024a] advocate for the design of "civic feedback processes that are robust against disparities in the representation of demographic and opinion minorities," including reweighting techniques "for more equitable voice among demographic minorities which were underrepresented in the process;" such reweighting could especially be appropriate for processes which are consultative for policymakers as opposed to binding.

Analogously, motivated by unequal volunteer and dropout rates in deliberative democracy, an important line of work shows how balanced panels can be selected ("sortition"). Both individual fairness (volunteers should have sufficiently high selection probabilities, even if from overrepresented groups) and overall representative balance (on both observed and unobserved covariates) are important [Benadè et al. 2019; Flanigan et al. 2020; Flanigan et al. 2021; Flanigan et al. 2021; Ebadian et al. 2022; Flanigan et al. 2023; Baharav and Flanigan 2024; Flanigan et al. 2024; Caragiannis et al. 2024; Ebadian and Micha 2025; Assos et al. 2025]. Their algorithms have been deployed at scale to support panel selection [Flanigan et al. 2021].

What lessons does this literature provide? (1) Representative participation is seen as a central design goal by researchers and practitioners, including in the EconCS community—it is well accepted that an unrepresentative process is not legitimate, though there is empirical equivocation on real-world participation disparities and its effect on resource allocation. (2) Representative participation (and overall rates) is nevertheless an ongoing challenge. In NYC participatory budgeting, fewer than 100,000 people vote city-wide,[2] far less than even other local elections such as for

[2]The exact number of people eligible is unclear, as it depends on the current number of residents



Variable	Coef	p-value
Intercept	−13720	< 0.001
Log Population	967	< 0.001
Log Median Income	472	0.021
Heat Vulnerability Index	−149	0.019

Fig. 1: In NYC, the number of tree planting requests by ZIP Code Tabulation Area in 2015-2024. Controlling for population, requests correlate positively with neighborhood income and *negatively* with a heat vulnerability index, a proxy for the need for shade. NYC no longer takes requests to plant trees, and instead will develop a schedule that prioritizes the most heat vulnerable areas.

city council—and some have questioned its use for decisionmaking [Golliher 2025]. It thus remains unclear how to mitigate the effects of heterogeneous participation.

2.2 Resident crowdsourcing

Another type of system deployed at scale to influence public resource allocation is *resident crowdsourcing*: people make service requests, such as through "311 systems" in the United States. NYC receives over 3 million requests a year—for incidents ranging from fallen trees on powerlines, to potholes, flooding, rodents, and to request new tree planting—and similar systems are in place in hundreds of cities globally. This is an important avenue for the government to learn about problems – supplementing and informing less frequent active inspections – and there is a large government bureaucracy to respond to requests.

However, substantial research, including my own, has established that participation is heterogeneous, even *conditional on ground truth conditions*. For example, in Liu et al. [2024], we show how to use *duplicate* reports about the same incident to estimate reporting delays; in Agostini et al. [2024], we use *spatial correlation* to probabilistically identify unreported incidents; in Balachandar et al. [2025], we combine regularly scheduled government inspections with crowdsourced reports; and in Franchi et al. [2025], we identify true flooding prevalence using a vision-language model on dashcams street imagery. In all cases, despite the diverse identification strategies, we find that (a) crowdsourced reporting data can be informative about ground truth conditions, e.g., that more hazardous conditions are reported at far higher rates [Liu et al. 2024]; but also (b) reporting is correlated with socioeconomic characteristics, also substantially: e.g., in Liu et al. [2024], we find that higher income, population density, voter participation, fraction of people with college degrees, and fraction of the population that is white all correlate with higher reporting rates. These patterns induce heterogeneous delays in incidents being

over 11 years old in the city council districts that participated. There are almost 8 million residents over 10 years old in NYC, implying a less than 2% turnout rate.

addressed, potentially leading to inequitable government service.

What should we do given this heterogeneous participation? I do *not* believe that these results imply we should not crowdsource this information; rather, we should ensure efficient resource allocation despite it, as in the work to balance citizen assemblies. For example, in a followup project, we seek to optimize inspection resources to efficiently and equitably set service level agreements [Liu and Garg 2024]; it would be conceptually simple to account for heterogeneous reporting delays.

I believe that the design of such modifications is urgent, before practitioners decide that information aggregation is not worth the resulting allocation skews. NYC no longer allows the public to request new tree planting locations; instead, the Department of Parks and Recreation will plant "street trees on a cyclical basis and prioritiz[e] the most heat-vulnerable neighborhoods first" [New York City 2025b]. A simple analysis using public data [New York City 2024; 2025a] helps explains why: as shown in fig. 1, planting requests historically correlated positively with median neighborhood income and *negatively* with heat vulnerability, one measure of "need." Optimizing solely for stated resident preferences would lead to inefficient allocation, when the government has some expertise. Such pullback may also occur in other settings, if the mechanisms are not viewed as legitimate.

2.3 School matching

Finally, consider school matching. In many urban environments, students are assigned to public schools through the deferred acceptance algorithm [Abdulkadiroğlu et al. 2005]. The algorithm inputs applicant preferences (via ranked lists of schools) and school priorities (with factors such as geography, academic performance, diversity, and lottery numbers). The promise is twofold: (a) these systems provide the opportunity to access desired schools, even if they are not in the student's neighborhood; (b) when slots in high-value schools are scarce, they are allocated not solely due to geography but also accounting for student preferences, academic performance, and random chance—thus, these allocation systems can be more effective and equitable than those that simply reflect geographic segregation.

In practice, applying effectively can be time consuming for families: in NYC, there are over 800 high school programs to choose from, each with varying locations, classes and sports teams offered, and school quality metrics. Families who can afford it often pay for admissions consultant services. A long line of research has empirically shown that information access, awareness, and the time-consuming process – not just preferences – affects application behavior, both in NYC high school admissions [Corradini 2024; Corradini and Idoux 2025] and elsewhere [Larroucau et al. 2024; Tomkins et al. 2023; Arteaga et al. 2022; Ajayi and Sidibe 2020].

The "administrative burdens" [Herd and Moynihan 2019] of applying lead to participation heterogeneity and outcome inequity. For example, Cohodes et al. [2022] documents the large fraction of students who apply to non-competitive, "nonoptimal" schools first in their rankings. In Peng et al. [2025], we show that such behavior leads to substantial "undermatching": students not matching to as high-quality programs as they could have (that are no further geographically than their actual match), because they did not apply. In particular, this gap between where students matched and where they could have matched is almost twice as large for the most competitive Black and Hispanic students as it is for the corresponding

Asian and white ones, when quality is measured by program performance, value add, selectivity, school graduation rate, or college enrollment rate. We then show that simple application behaviors explain a large portion of this undermatching. Including with surveys, Corradini and Idoux [2025] show that differential awareness of schools rated as high-quality and racial homophily preferences explain such gaps, as opposed to preferences over other characteristics like quality.

Substantial work has further gone into developing and evaluating *informational interventions* to close the participation gaps [Corcoran et al. 2018; Arteaga et al. 2022; Cohodes et al. 2022; Corradini 2024; Larroucau et al. 2024]—for example, by providing students lists of high-quality programs close to their neighborhood. These interventions have changed behavior, when used.

Despite this focus, much work remains to be done. Informational gaps and heterogeneous participation persist, as documented by recent studies [Corradini and Idoux 2025; Peng et al. 2025]. One constant challenge, highlighted by Cohodes et al. [2022], is that informational interventions only work to the extent that they are *used*, i.e., they put the burden on participants—just as targetted advertising for participatory budgeting ultimately requires people to respond to the ads. Thus, it remains open how to deploy (a) interventions with high takeup rates and (b) new mechanisms robust to heterogeneous participation.

2.4 Common Themes and Implications

The above examples all follow a similar pattern: a mechanism allocates scarce public resources or makes joint decisions; a core mechanism component is to input preferences or information from participants; when the mechanism is deployed, participation is heterogeneous, despite it being monetarily "free." Such heterogeneity both makes the mechanism less effective and potentially skews allocation and decisionmaking against those already disadvantaged. While substantial work has been done to measure and reduce these disparities, they persist.

Related concerns potentially apply in other settings in which preferences are elicited from participants who may have heterogeneous capabilities: in refugee matching, refugees may be asked for preferences over host countries [Jones and Teytelboym 2017]; in food bank allocation, large food banks (but not small ones) have dedicated staff to interface with the mechanism [Prendergast 2017]; in kidney exchange, preference elicitation from doctors regarding compatible kidneys is a practical challenge [Ashlagi and Roth 2021].

What should we do, given this fact pattern? In any given setting, the options are to (a) defend the status quo, by establishing that the mechanism is nevertheless effective, or at least preferable over any feasible counterfactual mechanism; (b) aim to reduce participation heterogeneity, as behavior is far from fixed; (c) reform the mechanism, so that it is robust; and/or (d) replace it entirely, likely to one that minimally uses the people's preferences and information. How should we choose which option(s) to pursue? Different applications have and should take different paths, and the paths are complementary.[3] In the face of heterogeneous participation, we

[3]There are key differences between the applications. Voting leads to a collective decision, and individuals who do not participate nevertheless benefit if they agree with those who do. In stable matching, allocations are individual and more arguably 'zero-sum.' Crowdsourcing lies in-between,

defend standard democracy *and* invest substantially in voter turnout efforts; this is also the path taken so far for participatory budgeting and deliberative democracy. In other cases, we've seen either reform or a retreat from participatory mechanisms.

I posit that reform should aim for the following "best of both worlds" north star: *use preferences and information from those who choose to participate, but provide a "sufficient" quality of service to those who do not.* In other words, we should attempt to retain the benefits of public participation while mitigating the resulting resource allocation skews. With this goal, we would still have "power users" who benefit from their invested time; however, non-participation would lead to a reasonable, default allocation. Of course, "sufficient" and "reasonable" are subjective, and themselves policy choices; when allocating scarce resources, these defaults may come at some cost to the "power users." Policymakers and the public are in the best position to choose the context-dependent operating point on the inevitable tradeoff between information aggregation and allocation skews. My position is that this choice should be explicit, as opposed to the too-common status quo of maximally supporting aggregation at the cost of allocation skews.

3. APPROACHES FOR PRACTICE

In the remainder of this article, I highlight potential directions for practice and research, in service of this goal. These paths are informed by the above literature and Herd and Moynihan [2019] in particular. They lay out three reforms to respond to the "Medicare Maze," in which the elderly must annually learn about complex options to choose a health care plan, leading to worse health outcomes and increased costs: (1) reduce choice by simplifying options; (2) expand outreach and human assistance in navigating the choices; (3) use administrative data and information technology to provide personalized defaults or recommendations. These options have their analogues for participatory mechanisms.

3.1 Reduce participation heterogeneity

The simplest response to participation heterogeneity is to try to reduce it. In participatory budgeting, this is done via targeted advertising; in deliberative democracy, this is done more directly by modifying selection probabilities.[4] This approach is also a key tool to reduce disparities in the takeup of other entitlements, like SNAP benefits in the United States; Koenecke et al. [2023] show public support for targeted advertising that improves allocation equity. However, as continued disparities prove, turnout efforts are not a panacea in the presence of structural barriers to participation, such as those discussed by Mullainathan and Shafir [2013]. Approaches that more directly tackle structural barriers, such as those that provide childcare and video conferencing technology for deliberative democracy, may be necessary.

Another, more systematic approach to reducing participation *heterogeneity* is to use preferences *within* areas with relatively homogeneous participation. For exam-

as allocations (e.g., pothole fixes) are geographically localized, but everyone may benefit from the information shared by participants. The exposition has ignored these differences, as they are not crucial to my core thesis. However, they may be relevant in considering paths forward.

[4]This is only possible because deliberative democracy purposely is designed to select a subset of the people, with the goal of making that subset representative.

ple, NYC runs participatory budgeting separately for each city council district, with a set budget per district; *if* districts are drawn such that participation is similar within each district, then heterogeneity across districts would not skew allocations.[5] Analogously, in the tree planting context, the following approach could incorporate geographic balance, need as determined by the agency, and resident requests: make neighborhood-level scheduling and quantity decisions according to agency expertise; then, within each neighborhood, allow requests to inform precise planting locations, alongside expertise. Appropriately designed, such an approach could be "best of both worlds" and combine elicited preferences with expert decisionmaking.

3.2 Provide personalized defaults or recommendations

Turnout campaigns may be effective when participation is (meaningfully approximated as) binary: in voting mechanisms and resident crowdsourcing, the most important outcomes are how regularly someone votes or submits requests. In school matching, on the other hand, *whether* people submit ranked lists is not the only concern, as doing so is required to enroll a child in public school. Rather, submitting *informed* ranked lists is a challenge, as it requires awareness of program quality and admissions probabilities; only some may have access to expensive consulting services or advice from social networks to help them navigate these decisions. In such systems, practitioners, alongside researchers, may be able to provide recommendations or even default options to users. Then, applicants can – just as in the status quo – provide preferences if they are dissatisfied with the recommendations or defaults; others can choose to follow the recommendations. Of course, as with targetted advertising, one challenge with recommendations is takeup [Cohodes et al. 2022], and so stronger user interfaces or "nudges" are important.

In many cases, as in school matching, there already *is* a default option, e.g., a manual administrative placement if an applicant does not match with any school. One approach is for these defaults to be more systemically planned, to provide better allocations to those who do (can) not participate meaningfully. Recommendations and defaults are also related to – and 'lighter-touch' than – another approach developed in school matching: limiting options, potentially in a data-driven manner: Shi [2015] develops short choice menus for each family in Boston, citing "too many options" as contributing to long commute times, unpredictability, loss of neighborhood cohesion, and a research burden on families; Allman et al. [2022] develop small zones in San Francisco, in support of school diversity.

The use of personalized defaults and recommendations, powered by modern machine learning methods, may also be effective in other contexts. Ashlagi and Roth [2021] advocate for a related approach in the context of preference elicitation difficulties for kidney exchanges: "it may be useful to develop machine learning models to predict positive crossmatches and ... to understand the trade-offs involved with waiting (while on dialysis) for a better match."

[5]It is not clear that NYC's districts meet this criteria. My district spans relatively wealthy areas in the Upper West Side, to Columbia University, to lower-income areas in West Harlem. However, granular turnout data is unavailable and winning projects did not geographically concentrate.

3.3 Actively acquire information or post-process inputs

A third approach is for the central system to actively invest in information acquisition to counter participation biases. When resident crowdsourcing informs resource allocation, for example, agencies can invest more active inspection resources or install sensors (e.g., flood sensors [Franchi et al. 2025]) in neighborhoods with lower reporting rates. Alternatively, given the public's heterogeneous inputs, the system can make decisions that are nevertheless balanced. As an example of such post-processing, consider our work with the New York Public Library on the holds system, which allows patrons to request books from any system branch to be sent to their local neighborhood branch; we first found that heterogeneous usage of the holds system (even conditional on overall library usage) led to a large net outflow of books from lower-income neighborhoods to higher-income ones [Liu et al. 2024]. We then designed a routing prioritization scheme between branches to mitigate such disparities [Liu et al. 2025], so that all holds requests could be fulfilled without disproportionately depleting branches in lower holds-use neighborhoods.

However, these approaches are not always feasible. In deferred acceptance, where applicant preferences are directly used, it is unclear where active information acquisition can be incorporated or how matches can be post-processed. In democratic systems such as participatory budgeting, weighting votes may conflict with other design principles, such as 'one-person-one-vote' (as discussed by [Gelauff and Goel 2024a]). Such approaches may be feasible when constructing error bars or using vote outcomes to advise final decision-makers; however, the question of "representativeness" (and *of whom*) remains, especially when participation correlates with unobserved features [Chasalow and Levy 2021].

These solutions are analogous to those proposed in algorithmic fairness, to counter disparities in prediction accuracy that are caused by heterogeneous unobserved confounding or missing data. There, data may be actively acquired or post-processed while using demographics as features [Chen et al. 2018; Noriega-Campero et al. 2019; Caton and Haas 2020; Cai et al. 2020; Garg et al. 2021; Liu and Garg 2021; Movva et al. 2023; Zink et al. 2024; Balachandar et al. 2024; Dong et al. 2025; Chiang et al. 2025]. There as well, post-processing may be infeasible, due to legal constraints or a general preference for "group-unaware" approaches (e.g., the recent affirmative action ban in college admissions in the United States, which also affect algorithms in the admissions process [Lee et al. 2024]). More generally, I believe that the goal of countering heterogeneous participation may further connect market design to algorithmic fairness, cf. Finocchiaro et al. [2021].

All three approaches use central resources to counter heterogeneous participation and pursue "best of both words": use elicited preferences, but mitigate allocation skews. Next, I discuss how researchers can contribute to the vision.

4. RESEARCH DIRECTIONS

Researchers have an important role to play in collaborating with practitioners on designing, deploying, and evaluating the above approaches. Researchers—including those who do not collaborate with practitioners—can also contribute in other ways. Below, I overview three approaches for a diverse range of skillsets: (a) empirically quantifying heterogeneous participation; (b) providing theoretical insight on

participation-allocation tradeoff and designing mechanisms to navigate it; (c) more directly considering human-computer-market interactions and interface design.

4.1 Empirically quantify heterogeneous participation

Academics – through open data, information requests, or practitioner collaboration – can help quantify participation heterogeneity. Methodologically, the challenge is that quantifying participation heterogeneity often requires disambiguating it from other, less concerning, explanations. In resident crowdsourcing, we must show that heterogeneous *conditions* cannot explain the discrepancy – that it is not the case that some neighborhoods report less because they encounter fewer incidents worth reporting [Liu et al. 2024; Agostini et al. 2024; Balachandar et al. 2025; Franchi et al. 2025]. In school matching, we must show that heterogeneous preferences – e.g., due to outside options or true heterogeneity in idiosyncratic preferences for certain schools or school characteristics – do not fully explain behavior, and that instead heterogeneous information plays an important role [Larroucau et al. 2024; Corradini and Idoux 2025; Corradini 2024]. This challenge often requires new statistical methods, analyzing natural experiments, or careful collection of "ground truth" data, such as through surveys and randomized controlled trials.

Such quantification helps provide an empirical underpinning with which interventions can be justified and well-engineered. For example, quantifying missing incident reports by neighborhood helps in the allocation of inspection and sensor resources in resident crowdsourcing, and quantifying heterogeneous awareness and behavior informs the design of personalized recommendations in school matching.

Finally, I note that empirically quantifying heterogeneous participation is related to two empirical lines of work: (1) preference estimation under strategic behavior [Agarwal and Somaini 2018; Calsamiglia et al. 2020], where the goal is to estimate preferences in non-strategyproof mechanisms, when (some) agents may be strategic; (2) empirical behavioral economics, that seeks to quantify how human behavior deviates from "optimal," including in strategyproof mechanisms. Here, my focus is on quantifying *heterogeneous* behavior and its effects on downstream resource allocation, especially when there is no formal cost or strategic incentive.

4.2 Theoretically model allocation under heterogeneous participation and design mechanisms to explicitly navigate the participation-allocation tradeoff

Theoretical modeling of heterogeneous participation is a rich area for further study, to complement empirical measurement. Models can (a) elucidate welfare outcomes under heterogeneous participation; and (b) help design better mechanisms.

In the context of school matching, Kloosterman and Troyan [2020] analyze a setting in which some students are more informed than others about high quality options; under the model, such students may be *worse off* under deferred acceptance than without school choice; they then show that priorities may be designed in a way to avoid this outcome. Pathak and Sönmez [2008] analyze matching settings in which some students are "sophisticated" (strategic), while others are sincere despite strategic incentives; while "sincere students lose priority to sophisticated students" under the non-strategy-proof Boston mechanism, "any sophisticated student weakly prefers her assignment under the Pareto-dominant Nash equilibrium of the Boston mechanism to her assignment under the recently adopted student-optimal

stable mechanism." It is essential to develop such models for other settings, as well as experiment with and deploy mechanisms with properties similar to the ones developed by Kloosterman and Troyan [2020]. More generally, some mechanisms may be more effective at supporting diverse participation.

One setting where such conceptual insights helped was the design of Feeding America's market mechanism to allocate food to food banks. As detailed by Prendergast [2017], an essential consideration was to protect smaller food banks from heterogeneous participation, as they have "fewer resources and manpower ... relative to their larger counterparts, where there are often dozens of workers or volunteers." The chosen mechanism avoided a continuous auction (which would benefit those with dedicated staff members) and allowed fractional bidding and storing of credits. It further effectively enabled a default option, giving food banks "the option to delegate bidding to an employee of Feeding America, where a food bank could simply outline in broad terms its needs to that person" [Prendergast 2017].

A related question suitable for modeling insight is: under what contexts is the participation-allocation tradeoff big, and when should we potentially abandon a mechanism? This question has recently been explored in the context of individual-level prediction to target resources: Shirali et al. [2024] argue that "prediction-based allocations outperform baseline methods using aggregate unit-level statistics only when between-unit inequality is low and the intervention budget is high," i.e., that the cost of individualized predictions may not be worth it; Perdomo et al. [2023] empirically illustrate such ideas in the context of targeting interventions for students at risk of dropping out of school. Wang et al. [2024] argue against the legitimacy of decisionmaking that uses predictions of the future about individuals, due to reoccurring challenges regarding accuracy, disparate performance, and related concerns. Analogously, it may emerge in a model that eliciting preferences is only worth it when heterogeneity from preferences is larger than that from participation.

4.3 More directly consider human-computer-market interaction

Finally, the EconCS community should increase collaborations with human-computer interaction (HCI) researchers, to build interfaces that more effectively allow equal participation. Schank and McGuinness [2021] and Pahlka [2023] both pinpoint bad interface design as worsening government service. I posit that (1) good interfaces may be more effective than good theoretical properties in improving participation and systems, and (2) qualitative studies are important to understand participation. Here, I briefly overview my work and collaborations with HCI researchers.

In Bartle et al. [2025], we build and deploy an SMS-based system to help place patients being discharged from hospitals into nursing care homes. In our context in Hawai'i, care homes are often run by retired nurses *out of their own homes*, with only one or two patients; whenever a patient needs to be placed, a full-time team of hospital social workers calls the approximately thousand nursing homes to see if they have capacity and can care for the given patient's needs. This preference information is not centrally available because integration into a healthcare management system like Epic does not work for this rural, single-operator population. As we show, simply collecting capacity information – through SMS – and showing the data to hospital social workers trying to place patients into homes improves the process; my conjecture is that this data improvement – enabled by effective interface design

for care home operators – is far greater than improved call recommendations, such as through a matching optimization, would yield. However, one challenge is that many homes do not share their preferences over patient characteristics with the system. In followup work, we ran a randomized controlled trial and interviewed care home operators to understand this participation gap [Bartle et al. 2025]. The experiment revealed that nudges can (somewhat) increase the number of homes who share their preferences, and the interviews uncovered complex cultural phenomena as well as economic considerations that shape dynamic preferences. The mixed methods approach and collaboration across fields was essential in understanding participation and how interventions may increase it.

Substantial work has also shown that interface design can affect behavior in other systems studied by the community. In participatory budgeting, substantial work compares the behavioral and learning implications of the elicitation mechanism (i.e., whether they voters asked to rank or approve projects, or to modify a proposed budget or give a full budget) [Gelauff et al. 2018; Garg et al. 2019; Garg et al. 2019; Goel et al. 2019; Gelauff and Goel 2024b]—with the hypothesis that some mechanisms may be easier for voters to understand. In ratings systems, simple modifications such as the question that is asked can substantially affect behavior, by aligning different people on what "five stars" actually means [Garg and Johari 2019; 2021]. Similarly, I conjecture that interface design could reduce participation gaps in other systems. While simplifying interfaces is likely to be generally useful, open questions remain on how to best present information, including recommendations. Future work should experimentally evaluate interfaces and qualitatively interview participants regarding how they perceive a given interface and system design.

5. CONCLUSION

Economic and computational researchers have important roles to play in designing and analyzing societal systems [Roth 2002; Abebe et al. 2020]. Our community should be proud of our impact in influencing the deployment of so many real-world systems. Undoubtedly, many of these systems improve upon those that they replaced. However, just as we theoretically design mechanisms to be strategyproof, so that people *can safely* share their true preferences, we should focus on whether people *do* participate on equal footing, or can do so in the presence of heterogeneous time costs. We should further engineer our systems – theoretically, algorithmically, and through interface design – so that they do not inadvertently allocate scarce resources according to participation ability. In this article, I overviewed research, including much of my own, in pursuit of this goal. I believe that "best of both worlds" systems, that incorporate preferences without allowing heterogeneous participation to skew distributional outcomes, are possible and necessary.

Acknowledgments

I thank Rediet Abebe, Bailey Flanigan, Lodewijk Gelauff, Paul Gölz, Allison Koenecke, Karen Levy, Irene Lo, and members of Cornell's AI, Policy, and Practice Initiative for invaluable discussion and feedback, as well as my advisors, students, and collaborators. NG is supported by NSF CAREER IIS-2339427, and Cornell Tech Urban Tech Hub, Meta, and Amazon research awards.

REFERENCES

Abdulkadiroğlu, A., Pathak, P. A., and Roth, A. E. 2005. The New York City High School Match. *American Economic Review 95,* 2 (Apr.), 364–367.

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D. G. 2020. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 252–260.

Agarwal, N. and Somaini, P. 2018. Demand analysis using strategic reports: An application to a school choice mechanism. *Econometrica 86,* 2, 391–444.

Agostini, G., Pierson, E., and Garg, N. 2024. A bayesian spatial model to correct under-reporting in urban crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 21888–21896.

Ajayi, K. and Sidibe, M. 2020. School choice under imperfect information. *Economic Research Initiatives at Duke (ERID) Working Paper* 294.

Allman, M., Ashlagi, I., Lo, I., Love, J., Mentzer, K., Ruiz-Setz, L., and O'Connell, H. 2022. Designing school choice for diversity in the san francisco unified school district. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. 290–291.

Arteaga, F., Kapor, A. J., Neilson, C. A., and Zimmerman, S. D. 2022. Smart matching platforms and heterogeneous beliefs in centralized school choice*. *The Quarterly Journal of Economics 137,* 3 (03), 1791–1848.

Ashlagi, I. and Roth, A. E. 2021. Kidney exchange: An operations perspective. *Management Science 67,* 9, 5455–5478.

Assos, A., Baharav, C., Flanigan, B., and Procaccia, A. 2025. Alternates, assemble! selecting optimal alternates for citizens' assemblies. *Available at SSRN 5283438*.

Baharav, C. and Flanigan, B. 2024. Fair, manipulation-robust, and transparent sortition. In *Proceedings of the 25th ACM Conference on Economics and Computation*. 756–775.

Balachandar, S., Garg, N., and Pierson, E. 2024. Domain constraints improve risk prediction when outcome data is missing. In *The Twelfth International Conference on Learning Representations*.

Balachandar, S., Sadhuka, S., Berger, B., Pierson, E., and Garg, N. 2025. Urban incident prediction with graph neural networks: Integrating government ratings and crowdsourced reports. *arXiv preprint arXiv:2506.08740*.

Bartle, V., Dell, N., and Garg, N. 2025. Shopping around: An experiment in preferences and incentives for placing long-term patients.

Bartle, V., Shearer, A., Wroe, A., Dell, N., and Garg, N. 2025. Faster information for effective long-term discharge: A field study in adult foster care. *Proceedings of the ACM on Human-Computer Interaction 9,* 2, 1–29.

Benadè, G., Gölz, P., and Procaccia, A. D. 2019. No stratification without representation. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 281–314.

Cai, W., Gaebler, J., Garg, N., and Goel, S. 2020. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 22–28.

Calsamiglia, C., Fu, C., and Guell, M. 2020. Structural estimation of a model of school choices: The boston mechanism versus its alternatives. *Journal of Political Economy 128,* 2 (Feb.), 642–680.

Caragiannis, I., Micha, E., and Peters, J. 2024. Can a few decide for many? the metric distortion of sortition. In *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. JMLR.org.

Caton, S. and Haas, C. 2020. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*.

Chasalow, K. and Levy, K. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 77–89.

Chen, I., Johansson, F. D., and Sontag, D. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems 31*.

Chiang, E., Shanmugan, D., Beecy, A. N., Sayer, G., Estrin, D., Garg, N., and Pierson, E. 2025. Learning Disease Progression Models That Capture Health Disparities. In *Conference on Health, Inference, and Learning (CHIL '25)*.

Cohodes, S. R., Corcoran, S. P., Jennings, J. L., and Sattin-Bajaj, C. 2022. When Do Informational Interventions Work? Experimental Evidence from New York City High School Choice. *Educational Evaluation and Policy Analysis*, 01623737231203293.

Corcoran, S. P., Jennings, J. L., Cohodes, S. R., and Sattin-Bajaj, C. 2018. Leveling the playing field for high school choice: Results from a field experiment of informational interventions. Working Paper 24471, National Bureau of Economic Research. March.

Corradini, V. 2024. Information and Access in School Choice Systems: Evidence from New York City.

Corradini, V. and Idoux, C. M. 2025. Overcoming racial gaps in school preferences: The effect of peer diversity on school choice. Tech. rep.

Dong, E., Schein, A., Wang, Y., and Garg, N. 2025. Addressing discretization-induced bias in demographic prediction. *PNAS nexus*, pgaf027.

Ebadian, S., Kehne, G., Micha, E., Procaccia, A. D., and Shah, N. 2022. Is sortition both representative and fair? *Advances in Neural Information Processing Systems 35*, 3431–3443.

Ebadian, S. and Micha, E. 2025. Boosting sortition via proportional representation. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '25. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 667–675.

Finocchiaro, J., Maio, R., Monachou, F., Patro, G. K., Raghavan, M., Stoica, A.-A., and Tsirtsis, S. 2021. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 489–503.

Fishkin, J., Garg, N., Gelauff, L., Goel, A., Munagala, K., Sakshuwong, S., Siu, A., and Yandamuri, S. 2019. Deliberative democracy with the online deliberation platform. In *Demo at the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019 Demo)*.

Fishkin, J. S. 1991. *Democracy and deliberation: New directions for democratic reform*. Yale University Press.

Flanigan, B., Gölz, P., Gupta, A., Hennig, B., and Procaccia, A. D. 2021. Fair algorithms for selecting citizens' assemblies. *Nature 596,* 7873, 548–552.

Flanigan, B., Gölz, P., Gupta, A., and Procaccia, A. D. 2020. Neutralizing self-selection bias in sampling for sortition. *Advances in Neural Information Processing Systems 33*, 6528–6539.

Flanigan, B., Gölz, P., and Procaccia, A. 2023. *Mini-Public Selection: Ask What Randomness Can Do for You*. Ash Institute for Democratic Governance and Innovation.

Flanigan, B., Kehne, G., and Procaccia, A. D. 2021. Fair sortition made transparent. 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., 25720–25731.

Flanigan, B., Liang, J., Procaccia, A. D., and Wang, S. 2024. Manipulation-robust selection of citizens' assemblies. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press.

Franchi, M., Garg, N., Ju, W., and Pierson, E. 2025. Bayesian modeling of zero-shot classifications for urban flood detection. *arXiv preprint arXiv:2503.14754*.

Garg, N., Gelauff, L. L., Sakshuwong, S., and Goel, A. 2019. Who is in your top three? optimizing learning in elections with many candidates. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 22–31.

Garg, N. and Johari, R. 2019. Designing optimal binary rating systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1930–1939.

Garg, N. and Johari, R. 2021. Designing informative rating systems: Evidence from an online labor market. *Manufacturing & Service Operations Management 23,* 3, 589–605.

Garg, N., Kamble, V., Goel, A., Marn, D., and Munagala, K. 2019. Iterative local voting for collective decision-making in continuous spaces. *Journal of Artificial Intelligence Research 64*, 315–355.

Garg, N., Li, H., and Monachou, F. 2021. Standardized tests and affirmative action: The role of bias and variance. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 261–261.

Gelauff, L. and Goel, A. 2024a. Opinion change or differential turnout: Changing opinions on the austin police department in a budget feedback process. *Digit. Gov.: Res. Pract. 5,* 3 (Sept.).

Gelauff, L. and Goel, A. 2024b. Rank, pack, or approve: Voting methods in participatory budgeting. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 18. 448–461.

Gelauff, L., Goel, A., Munagala, K., and Yandamuri, S. 2020. Advertising for demographically fair outcomes. *arXiv preprint arXiv:2006.03983*.

Gelauff, L., Sakshuwong, S., Garg, N., and Goel, A. 2018. Comparing voting methods for budget decisions on the assu ballot. Tech. rep., Technical report.

Goel, A., Krishnaswamy, A. K., Sakshuwong, S., and Aitamurto, T. 2019. Knapsack voting for participatory budgeting. *ACM Transactions on Economics and Computation (TEAC) 7,* 2, 1–27.

Golliher, D. 2025. The city council's participatory budgeting by the numbers. Maximum New York.

Hartig, H., Daniller, A., Keeter, S., and Van Green, T. 2023. Republican gains in 2022 midterms driven mostly by turnout advantage.

Hayduk, R., Hackett, K., and Folla, D. T. 2017. Immigrant engagement in participatory budgeting in new york city. *New Political Science 39,* 1, 76–94.

Herd, P. and Moynihan, D. P. 2019. *Administrative burden: Policymaking by other means*. Russell Sage Foundation.

Johnson, C., Carlson, H. J., and Reynolds, S. 2023. Testing the participation hypothesis: Evidence from participatory budgeting. *Political Behavior 45,* 1, 3–32.

Jones, W. and Teytelboym, A. 2017. The international refugee match: A system that respects refugees' preferences and the priorities of states. *Refugee Survey Quarterly 36,* 2, 84–109.

Kloosterman, A. and Troyan, P. 2020. School choice with asymmetric information: Priority design and the curse of acceptance. *Theoretical Economics 15,* 3, 1095–1133.

Koenecke, A., Giannella, E., Willer, R., and Goel, S. 2023. Popular support for balancing equity and efficiency in resource allocation: A case study in online advertising to increase welfare program awareness. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 17. 494–506.

Larroucau, T., Rios, I., Fabre, A., and Neilson, C. 2024. College application mistakes and the design of information policies at scale.

Lee, J., Harvey, E., Zhou, J., Garg, N., Joachims, T., and Kizilcec, R. F. 2024. Ending affirmative action harms diversity without improving academic merit. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–17.

Lee, S. 2024. Deliberative polling on constitutional amendments in mongolia. Tech. rep., Global Assembly.

Liu, Z., Bhandaram, U., and Garg, N. 2024. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science 4,* 1, 57–65.

Liu, Z. and Garg, N. 2021. Test-optional policies: Overcoming strategic behavior and informational gaps. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–13.

Liu, Z. and Garg, N. 2024. Redesigning service level agreements: Equity and efficiency in city government operations. In *Proceedings of the 25th ACM Conference on Economics and Computation*. 309–309.

Liu, Z., Rankin, S., and Garg, N. 2024. Identifying and addressing disparities in public libraries with bayesian latent variable modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 22258–22265.

Liu, Z., Zhu, W., Rankin, S., and Garg, N. 2025. Optimizing library usage and browser experience: Application to the new york public library. *arXiv preprint arXiv:2503.23118*.

Movva, R., Shanmugam, D., Hou, K., Pathak, P., Guttag, J., Garg, N., and Pierson, E. 2023. Coarse race data conceals disparities in clinical risk score performance. In *Machine Learning for Healthcare Conference*. PMLR, 443–472.

Mullainathan, S. and Shafir, E. 2013. *Scarcity: Why having too little means so much*. Macmillan.

New York City. 2024. Heat vulnerability index rankings. Available at: `https://data.cityofnewyork.us/Health/Heat-Vulnerability-Index-Rankings/4mhf-duep`; Accessed: May 16, 2025.

New York City. 2025a. 311 service requests from 2010 to present. Available at: `https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9`; Accessed: May 16, 2025.

New York City. 2025b. Tree planting. Available at: `https://portal.311.nyc.gov/article/?kanumber=KA-01895`, Accessed: May 16, 2025.

Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., and Pentland, A. 2019. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 77–83.

Pahlka, J. 2023. *Recoding America: why government is failing in the digital age and how we can do better*. Metropolitan Books.

Pathak, P. A. and Sönmez, T. 2008. Leveling the playing field: Sincere and sophisticated players in the boston mechanism. *American Economic Review 98,* 4, 1636–1652.

Peng, K., Ryu, E., Kleinberg, J., Tardos, E., and Garg, N. 2025. Deviations from reach-match-safety strategies explain undermatching disparities in new york city high schools.

Perdomo, J. C., Britton, T., Hardt, M., and Abebe, R. 2023. Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205*.

Prendergast, C. 2017. How food banks use markets to feed the poor. *Journal of Economic Perspectives 31,* 4, 145–162.

Roth, A. E. 2002. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica 70,* 4, 1341–1378.

Schank, H. and McGuinness, T. D. 2021. *Power to the public: The promise of public interest technology*. Princeton University Press.

Shen, Z., Gelauff, L., Goel, A., Korolova, A., and Munagala, K. 2021. Robust allocations with diversity constraints. *Advances in Neural Information Processing Systems 34*.

Shi, P. 2015. Guiding school-choice reform through novel applications of operations research. *Interfaces 45,* 2, 117–132.

Shirali, A., Abebe, R., and Hardt, M. 2024. Allocation requires prediction only if inequality is low. *arXiv preprint arXiv:2406.13882*.

Shybalkina, I. and Bifulco, R. 2019. Does participatory budgeting change the share of public funding to low income neighborhoods? *Public Budgeting & Finance 39,* 1, 45–66.

Stewart, L. M., Miller, S. A., Hildreth, R., and Wright-Phillips, M. V. 2014. Participatory budgeting in the united states: a preliminary analysis of chicago's 49th ward experiment. *New Political Science 36,* 2, 193–218.

Tomkins, S., Grossman, J., Page, L., and Goel, S. 2023. Showing high-achieving college applicants past admissions outcomes increases undermatching. *Proceedings of the National Academy of Sciences 120,* 45, e2306017120.

Wang, A., Kapoor, S., Barocas, S., and Narayanan, A. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing 1,* 1, 1–45.

Zepic, R., Dapp, M., and Krcmar, H. 2017. Participatory budgeting without participants: Identifying barriers on accessibility and usage of german participatory budgeting. In *2017 Conference for E-Democracy and Open Government (CeDEM)*. IEEE, 26–35.

Zink, A., Obermeyer, Z., and Pierson, E. 2024. Race adjustments in clinical algorithms can help correct for racial disparities in data quality. *Proceedings of the National Academy of Sciences 121,* 34, e2402267121.

Calibration through the Lens of Indistinguishability

PARIKSHIT GOPALAN
Apple
and
LUNJIA HU
Harvard University

Calibration is a classical notion from the forecasting literature which aims to address the question: how should predicted probabilities be interpreted? In a world where we only get to observe (discrete) outcomes, how should we evaluate a predictor that hypothesizes (continuous) probabilities over possible outcomes? The study of calibration has seen a surge of recent interest, given the ubiquity of probabilistic predictions in machine learning. This survey describes recent work on the foundational questions of how to define and measure calibration error, and what these measures mean for downstream decision makers who wish to use the predictions to make decisions. A unifying viewpoint that emerges is that of calibration as a form of indistinguishability, between the world hypothesized by the predictor and the real world (governed by nature or the Bayes optimal predictor). In this view, various calibration measures quantify the extent to which the two worlds can be told apart by certain classes of distinguishers or statistical measures.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms: Measurement, Reliability, Theory

Additional Key Words and Phrases: Calibration, Uncertainty quantification, Prediction, Decision making

1. INTRODUCTION

Prediction is arguably the ubiquitous computational task of our time. Every day, a remarkable amount of computational resources are invested in the prediction of various probabilities, whether it is a language model trying to answer a user's ambiguous query or a recommendation engine trying to predict which product/profile to show a user. These automated predictions affect nearly every aspect of our lives, be it social, medical or financial. What makes prediction different from more classical computational tasks (such as sorting numbers or computing max-flows) is that there is no well-defined notion of what constitutes correctness.

To explore this issue in greater detail, let us consider the simplified setting of binary prediction, where nature is modeled as a joint distribution $\mathcal{D}^*$ over attributes $\mathbf{x}$ drawn from a domain $\mathcal{X}$ and labels $\mathbf{y} \in \mathcal{Y}$. In this article, we will mainly focus on the setting $\mathcal{Y} = \{0, 1\}$ of Boolean labels.[1] We denote the marginal distribution over $\mathcal{X}$ by $\mathcal{D}^*_{\mathcal{X}}$, and $\mathcal{Y}$ by $\mathcal{D}^*_{\mathcal{Y}}$. A predictor is a function $p : \mathcal{X} \to [0, 1]$. The *ground truth* in this setting is represented by the Bayes optimal predictor $p^*(x) = \mathbb{E}[\mathbf{y}^*|\mathbf{x} = x]$.

[1]We use boldface for random variables, thus $\mathbf{x}$ is random variable drawn from $\mathcal{X}$ whereas $x \in \mathcal{X}$ is a point in the domain.

Authors' addresses: parikg@apple.com, lunjia@alumni.stanford.edu

The obvious formulation of correctness in prediction might be to learn p^*. The challenge is that we never see the values of p^* itself, our only access to it is via the labels $\mathbf{y}^*$ which satisfy $\mathbb{E}[\mathbf{y}^*|x] = p^*(x)$. So the obvious formulation of correctness, as finding p which is close to p^* under some suitable measure of distance, will not work. There are (at least) two different and complementary approaches: loss minimization and calibration.

1.1 Loss minimization

In loss minimization, we choose a loss function $\ell : \{0,1\} \times [0,1] \to \mathbb{R}$, and a hypothesis class of predictors $\mathcal{P} = \{p : \mathcal{X} \to [0,1]\}$, and aim to find the predictor that minimizes

$$p = \arg\min_{p' \in \mathcal{C}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}^*} [\ell(\mathbf{y}^*, p'(\mathbf{x}))].$$

In essence, we use the labels $\mathbf{y}^*$ as a proxy for p^*, while ℓ plays the role of a distance measure. But it turns out that (for any proper loss) we indeed find the predictor in our family $\mathcal{P}$ that is closest to p^*. This is a consequence of the *bias variance* decomposition. Taking the example of squared loss $\ell(y,v) = (y-v)^2$, the decomposition tells us that for any predictor p, [2]

$$\mathop{\mathbb{E}}_{\mathcal{D}^*} [(\mathbf{y}^* - p(\mathbf{x}))^2] = \underbrace{\mathop{\mathbb{E}}_{\mathcal{D}^*} [(p(\mathbf{x}) - p^*(\mathbf{x}))^2]}_{\text{bias}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{D}^*} [(\mathbf{y}^* - p^*(\mathbf{x}))^2]}_{\text{variance}}.$$

Note that the variance is a property of p^*, independent of p.

Loss minimization is a simple yet immensely powerful paradigm that powers much of contemporary machine learning. But is it a satisfactory notion of correctness for prediction tasks? Here are some questions to consider:

—Imagine that a decision maker is using a predictor to make decisions that minimize their own loss function. This loss may differ from the one used to train the model, and might differ across various decision makers. For instance, we could use forecasts about rain to decide whether or not to carry an umbrella, to decide whether to have a party outdoors or indoors, or whether to turn off the sprinklers. Each of these has its own loss function. Say our loss for carrying an umbrella when it does not rain is 0.1, and for not carrying an umbrella when it rains is 0.9. The optimal strategy here is to carry an umbrella on days when $p^*(x) \geq 0.1$. Now suppose that the predictor p we have access to is not Bayes optimal. How do we make decisions using this predictor? Should we carry an umbrella whenever $p(x) \geq 0.1$, just like with the Bayes optimal predictor, or should we make decisions differently?

—We know that the squared loss decomposes into bias and variance, but we have no way of knowing how large each of these are. If we suffer large squared loss, it could because nature is inherently random (e.g p^* is often close to $1/2$), or because nature is deterministic but sufficiently complex that it *looks random* to

[2] It is easy to prove a similar statement about any *proper* loss, and a little harder to prove it about arbitrary losses. But the takeaway remains the same: by minimizing loss over a family $\mathcal{P}$, we find the best approximation to p^* from $\mathcal{P}$ under a suitable notion of distance tailored to the loss.

our hypothesis class $\mathcal{P}$. Loss minimization does not distinguish between these scenarios.

—Suppose we wish to predict the probability of rain tomorrow, and the model p found by minimizing squared loss gives a 60% chance of rain. How should we interpret this prediction? Is it possible that although p minimizes expected error globally, it is not particularly good at prediction for certain types of days (like days in September)? Concerns like these arise naturally in the context of fair predictions for subgroups (see the discussion on multicalibration in Section 1.4).

The question of what a prediction really guarantees naturally leads us to calibration.

1.2 Calibration

Calibration is a notion of correctness that focuses on ensuring that predicted probabilities align with actual outcomes. Intuitively, on days when a calibrated predictor predicts a 60% chance of rain, it rains 60% of the time. Formally, we can define perfect calibration as follows:

Definition 1.1. The predictor $p : \mathcal{X} \to [0,1]$ is *perfectly calibrated* under the distribution $\mathcal{D}^*$ if for every $v \in \mathsf{Image}(p)$, it holds that $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x}) = v] = v$.

A key property of calibration is that it simplifies downstream decision making. For instance, let us return to the problem of using the forecast about rain to decide whether or not to carry an umbrella, where our loss for carrying an umbrella when it does not rain is 0.1, and for not carrying an umbrella when it rains is 0.9. Now suppose that the predictor p we have access to is not Bayes optimal, but it is calibrated. If we are basing decisions solely on p, then the optimal strategy is still to carry an umbrella on days when $p(x) \geq 0.1$. The expected loss we would suffer is exactly what we would have suffered if our predictor were Bayes optimal.

This naturally motivates an alternate view of calibration as a notion of correctness for predictors based on indistinguishability from the Bayes optimal, which will be an important theme in this survey. This view is inspired by the outcome indistinguishability framework of [Dwork et al. 2021].[3]

To every predictor $p : \mathcal{X} \to [0,1]$, we can associate a distribution $\mathcal{D}^p$ on pairs $(\mathbf{x}, \mathbf{y}^p)$ where the marginal on $\mathbf{x}$ is $\mathcal{D}^*_{\mathcal{X}}$ and where $\mathbb{E}[\mathbf{y}^p|\mathbf{x}] = p(\mathbf{x})$. The Bayes optimal predictor for $\mathcal{D}^p$ is p. Perfect Calibration requires that the joint distributions $(p(\mathbf{x}), \mathbf{y}^p)$ and $(p(\mathbf{x}), \mathbf{y}^*)$ be identical.

Lemma 1.2 Perfect calibration as indistinguishability. *The predictor $p : \mathcal{X} \to [0,1]$ is perfectly calibrated under the distribution $\mathcal{D}^*$ iff the joint distributions $J^* = ((p(\mathbf{x}), \mathbf{y}^*))$ and $J^p = (p(\mathbf{x}), \mathbf{y}^p)$ on $[0,1] \times \{0,1\}$ are identical.*

Let us see why this is true. Since the marginal distribution of $\mathbf{x}$ is the same in both cases, the distribution of $p(\mathbf{x})$ is also the same. In essence, we require that the distributions $\mathbf{y}^*|p(\mathbf{x})$ and $\mathbf{y}^p|p(\mathbf{x})$ be identical. Since the latter is the Bernoulli distribution with parameter $p(\mathbf{x})$, we require the same for $\mathbf{y}^*|p(\mathbf{x})$, which

[3]That work does not consider calibration *per se*, it instead considers more general notions such as multicalibration from [Hébert-Johnson et al. 2018]. In the context of calibration, it is plausible that this indistinguishability viewpoint predates it, though we have not found a reference.

is the standard definition. This guarantee conditional on each prediction is the key strength of calibration as a prediction guarantee.[4]

The indistinguishability property asserts that $p(\mathbf{x})$ be a plausible explanation for the observations $\mathbf{y}^*$ given $\mathbf{x}$, in that the conditional distribution of $\mathbf{y}^*|p(\mathbf{x})$ is consistent with the hypothesis that $p(\mathbf{x})$ is the Bayes optimal predictor. This indistinguishability property is desirable in machine learning, where we often try to model complicated processes (like the likelihood of a medical condition) and are unlikely to find the true Bayes optimal. Calibrated predictors are considered more trustworthy, whereas a predictor that is not calibrated will fail some basic tests: the probability of the label being 1 conditioned on $p(\mathbf{x}) = v$ is not v.

In the bigger picture, the notion of indistinguishability has played a central role in several disciplines within theoretical computer science, cryptography and pseudorandomness to name just a couple, indeed its roots go back to the Turing test. Viewing calibration as a form of indistinguishability lets us draw on ideas from those areas when we seek to define approximate calibration or generalize our notions beyond the binary classification setting.

1.3 From perfect to approximate calibration

Perfect calibration is a clean abstraction, but predictors trained and used for prediction tasks in the real world are seldom perfectly calibrated. For calibration to be a useful notion, we need to define what it means for a predictor to be approximately (but not perfectly) calibrated, and we need efficient methods to measure calibration error. How to do this in a principled manner is the main focus of this article.

There are many desiderata that one might hope a notion of approximate calibration satisfies:

(1) It should preserve the desirable properties of calibration, such as indistinguishability and simple downstream decision making, in some approximate sense.

(2) It should be efficient to measure the calibration error of a given predictor, just from black box access to samples of the form $(p(\mathbf{x}), \mathbf{y}^*)$, both in terms of sample complexity and computational complexity. In an online setting (to be defined shortly), we might wish for our notion to have low-regret algorithms.

(3) The notion should be robust to small perturbations in the predictor. A tiny change to a calibrated predictor should not result in a predictor with huge calibration error. For instance, changing the days forecast from 60% to 59.999% should not result in wild swings in the calibration error.[5]

(4) The notion should extend beyond binary classification, to multiclass labeling and regression, while maintaining properties like efficiency.

Achieving all of these properties is not easy. The classical notion of calibration error, which is the expected calibration error or ECE, only satisfies property (1) above; we will discuss this in more detail in Section 3. An active line of recent

[4]Of course, there might be a different calibrated predictor that only puts the chances of rain at 30%. There is no contradiction because the level sets of the predictors over which we average are different in the two cases.

[5]This is especially desirable from a machine learning perspective, where the lower order bits of prediction are considered insignificant and typically disregarded in low-precision arithmetic.

research has yielded a rich theory of approximate notions of calibration, together with algorithms for computing them efficiently in various models. Yet, to date, there is no single notion that satisfies all four desiderata mentioned above!

Perhaps this is too much to hope for, since some of these desiderata (eg. robustness and low-regret algorithms) arise from different motivating scenarios. But a clear takeaway from this body of research is that approximate calibration is surprisingly challenging to define and measure. The key technicality in defining approximate calibration error comes from conditioning. Every definition of calibration involves some form of conditioning on predictions. While this conditioning is simple for perfectly calibrated predictors, it is far trickier for predictors that are not perfectly calibrated, since predictions are real-valued.

In this survey, we will highlight how the *indistinguishabilty* viewpoint on calibration guides us in formulating what approximate calibration should mean. At a high level, there are two approaches to this task:

—**Limit the set of distinguishers :** Rather than require J^* and J^p be identical, we ask that they look similar to a family W of distinguishers. The calibration error is measured by the maximum distinguishing advantage achieved over all distinguishers in W. This approach is directly inspired by cryptography and pseudorandomness.

—**Use a divergence/distance on distributions:** Since J^* and J^p are both distributions on the domain $[0,1]\times\{0,1\}$, we can use distance measures/divergences on probability distributions (e.g., total variation, earthmover) to measure the distance between them, and use this as our measure of the calibration error. As we will see, this view relates to a quantification of the economic value of calibration from the perspective of downstream decision making.

These approaches lead to a number of calibration error measures that we will explore in more detail in this article, and which have many advantages over ECE and other traditional calibration measures. We will analyze smooth calibration error [Kakade and Foster 2008], which satisfies properties (1-3) but not (4). It also corresponds to an intuitive notion of approximate calibration, where the predictor is close to some perfectly calibrated predictor in earthmover distance.

From the computational standpoint, the natural model in which to study calibration has been the online setting, where we measure the regret or calibration error of our prediction strategy over T time steps.[6] The classic work of [Foster and Vohra 1998] showed that sublinear calibration error, as measured by ECE is possible. The regret rate achieved in their work is $O(T^{2/3})$.[7] It is known that regret rates of $O(\sqrt{T})$ or even $\widetilde{O}(\sqrt{T})$ are not possible for ECE [Qiao and Valiant 2021], and figuring out the optimal regret achievable is an active area of research (see, e.g., [Dagan et al. 2025]). However, new notions of calibration, which we will discuss in this survey, actually admit prediction strategies that achieve $O(\sqrt{T})$ or $\widetilde{O}(\sqrt{T})$ regret rates [Qiao and Zheng 2024; Arunachaleswaran et al. 2025; Hu and Wu 2024].

[6]Note that the computational task of learning a calibrated predictor admits trivial solutions in the offline model; for instance, one can always predict the expectation of the label.

[7]The regret rate is T times the calibration error on the uniform distribution over the T time steps.

1.4 Limitations and generalizations

Calibration is clearly a desirable property for a predictor, but it has limitations, and cannot be considered as a standalone notion of goodness for a predictor. We ideally want predictors to have both good calibration and other properties like small expected loss. We discuss these limitations below, and use this as motivation to introduce the stronger notion of multicalibration [Hébert-Johnson et al. 2018], and discuss how it addresses these limitations.

Calibration does not guarantee utility. There are many predictors that will satisfy calibration, and we would not consider all of them to be equally informative or good. For instance, the *average* predictor $\bar{p}$ that always predicts the average label $\mathbb{E}_{\mathcal{D}^*}[\mathbf{y}^*]$ is perfectly calibrated, as is the Bayes optimal p^*. Any reasonable loss function would distinguish between these predictors, but calibration (by itself) does not.

Calibration gives guarantees on average over the entire population. In some applications, this might not be good enough. For instance, suppose we train a predictor to predict the risk of a certain risk of disease for a patient. On examining the data, we find that although the predictor is calibrated over the general population, it is miscalibrated for patients with a certain medical history, who are a small fraction of the dataset (so this does not affect the overall calibration error too much). We would not trust such a predictor to make decisions for those patients.

Multicalibration. Multicalibration, introduced in [Hébert-Johnson et al. 2018], is a strengthening of calibration. It requires that our predictions are calibrated, even when conditioned on membership in a rich collection of demographic subgroups $\mathcal{C} \subseteq 2^{\mathcal{X}}$. Which subgroups to consider is an important consideration, which is dictated by the data and computational resources available to the predictor. We refer the reader to [Hébert-Johnson et al. 2018] for more details.

Although calibration by itself does not guarantee good loss minimization, multicalibration with respect to rich class of subgroups $\mathcal{C}$ does imply strong loss minimization. This was the key insight in the work of [Gopalan et al. 2022] which introduced the notion of omniprediction. Omniprediction asks for a predictor that is as good as benchmark class $\mathcal{C}$ not just for a single loss function, but for any loss from a large family of loss functions. [Gopalan et al. 2022] shows a surprising connection between omniprediction with respect to a benchmark class $\mathcal{C}$ and multicalibration with respect to $\mathcal{C}$.

From the indistinguishability perspective, [Dwork et al. 2021] showed that multicalibration is equivalent to indistinguishability of the distributions $(c(\mathbf{x}), p(\mathbf{x}), \mathbf{y}^*)$ and $(c(\mathbf{x}), p(\mathbf{x}), \mathbf{y}^p)$ for all $c : \mathcal{X} \to \{0, 1\}$ that lie in some family $\mathcal{C}$ of functions. Beyond its original motivation in multigroup fairness, multicalibration has proved to be tremendously powerful, finding applications to omniprediction [Gopalan et al. 2022], domain adaptation [Kim et al. 2022], pseudorandomness [Dwork et al. 2023], and computational complexity [Casacuberta et al. 2024].

Organization of this survey. In Section 2, we consider expected calibration error (ECE) and explore its weaknesses. In Section 3, we introduce weighted calibration measures which capture the notion of indistinguishability to limited classes of distinguishers. This unifies several different notions of approximate calibration

in the literature. In Section 4, we describe Calibration decision loss, which looks at calibration from an economics perspective, through the eyes of a downstream decision maker who wants to use the predictions of a predictor to optimize their utility. We review the active area of research on online calibration in Section 5. Given the number of calibration notions that we will encounter, a natural question is whether there is some ground truth notion against which we can compare these different notions. In Section 6, we define the distance to calibration, which proposes a ground-truth notion of what approximate calibration ought to mean, and show how smooth calibration shows up naturally in this setting. In the interest of brevity, we omit most proofs from the survey. We direct the interested reader to the arXiv for a fuller version of this article that includes full proofs and some additional material.

2. EXPECTED CALIBRATION ERROR

We start with what is arguably the most popular metric for measuring calibration error: the expected calibration error or ECE. We examine some of its shortcomings, which will guide us in formulating other notions of approximate calibration.

Definition 2.1. The *expected calibration error* of a predictor p under $\mathcal{D}^*$ is defined as $\mathrm{ECE}(p, \mathcal{D}^*) = \mathbb{E}\,|\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] - p(\mathbf{x})]|$.

Some notes on the definition of ECE:

—While perfect calibration requires $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] = p(\mathbf{x})$, ECE allows for some slack in the equality, and measures the average deviation over all p.

—We have defined ECE as measuring the absolute deviation between $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})]$ and $p(\mathbf{x})$. We could instead have used the square or the q^{th} power for $q \geq 1$ and defined $\mathrm{ECE}_q(p, \mathcal{D}^*) = \mathbb{E}[|\,\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] - p(\mathbf{x})|^q]^{1/q}$. By the convexity of t^q, ECE_q is an increasing function of q.

For a better understanding of ECE, we look at two alternative characterizations. The first characterizes it in terms of the maximum inner product with a distinguisher b which is a bounded function on $[0, 1]$.

LEMMA 2.2. *Let* $B = \{b : \{0, 1\} \to [-1, 1]\}$ *be the family of all bounded functions. Then* $\mathrm{ECE}(p, \mathcal{D}^*) = \max_{b \in B} \mathbb{E}_{J^*}[b(x)(\mathbf{y}^* - p(\mathbf{x}))]$.

For two distributions $\mathcal{D}_1, \mathcal{D}_2$ on a domain $\mathcal{X}$, we define

$$\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) = \max_{S \subseteq \mathcal{X}} |\mathcal{D}_1(S) - \mathcal{D}_2(S)|.^8$$

We state the second characterization in terms of total variation distance.

LEMMA 2.3. *We have* $\mathrm{ECE}(p, \mathcal{D}^*) = \mathrm{TV}(J^*, J^p)$.

The trouble with ECE. At first glance, ECE seems to be a reasonable measure of calibration error. However there are (at least) a couple of problems with it: it is hard to efficiently estimate (even in the binary classification setting), and it is very discontinuous. Thus it fails desiderata (2-4).

[8]When the space $\mathcal{X}$ is infinite, we must restrict S to be measureable, but we will ignore this and other such subtleties.

The computational difficulty stems from Lemma 2.2. Estimating the ECE is equivalent to finding the best witness $b \in B$. This is essentially a learning problem over a class with infinite VC dimension. Indeed, one can show that sample complexity of estimating ECE can be as large as $\Omega(\sqrt{|\mathcal{X}|})$. Ideally, we would like to complexity to be independent of the domain size, and depending only on the desired estimation error.

The continuity problems are hinted at by Lemma 2.3. While total variation distance is a good distance measure for distributions over discrete domains, it is not ideal for continuous domains. And our setting involving distributions over predictions in $[0, 1]$ is inherently continuous. As the next example illustrates, ECE turns out to be highly discontinuous in the predictions of our predictor.

—Let $\mathcal{D}_2$ be the uniform distribution a two point space $\{(a, 0), (b, 1)\}$, where a is always labeled 0 and b is labeled 1.

—Consider the predictor p_0 which predicts $1/2$ for both a and b. It is perfectly calibrated, hence $\mathrm{ECE}(p_0) = 0$.

—For $\epsilon > 0$, define the predictor p_ϵ where $p_\epsilon(a) = 1/2 - \epsilon, p_\epsilon(b) = 1/2 + \epsilon$. It is easy to verify that $\mathrm{ECE}(p_\epsilon) = 1/2 - \epsilon$.

Think of ϵ being infinitesimally small but positive, so that p_ϵ is extremely close to p_0. Intuitively, p_ϵ is very close to being perfectly calibrated, it only requires a small perturbation of the lower order bits. Yet, the ECE is close to $1/2$ for p_ϵ, whereas it is 0 for p_0.

There are many ad-hoc fixes in practice that aim to get around these difficulties. For instance, bucketed ECE divides the interval $[0, 1]$ into b equal sized buckets, rounds the predictions in each bucket (say to the midpoint) and then measures the ECE of the discretized predictor. But [Blasiok et al. 2023a] observe that this results in a bucketed ECE which oscillates between 0 and $1/2 - \epsilon$ depending on whether the number of buckets is odd or even!

Are our issues with ECE small technicalities or symptoms of a bigger problem? We believe it is the latter. Assume you are training a predictive model, and you measure its ECE and find it to be large. Is this something you should worry about? Is your model truly miscalibrated (whatever that means)? Or is there an infinitesimal perturbation of its predictions that will make it perfectly calibrated? In general, there are sound reasons to prefer metrics that are reasonably smooth. It is also important for estimation to be efficient in terms of both samples and computation, which is not the case for ECE.

3. WEIGHTED CALIBRATION ERROR

In this section, we will explore notions of approximate calibration that only require that J^* and J^p look similar to a family W of distinguishers or weight functions. This results in a general template called weighted calibration, which is parametrized by the family W. Instantiating this notion with the family of bounded Lipschitz functions, we derive the notion of smooth calibration [Kakade and Foster 2008]. We briefly describe some other notions of calibration from the literature that can be viewed as instantiations of this template.

3.1 Weighted calibration

Weighted calibration error [Gopalan et al. 2022] captures the extent to which a collection of distinguishing functions are able to distinguish J^* from J^p. Since J^* and J^p are both distributions over $[0,1] \times \{0,1\}$, we consider distinguishing functions $f : [0,1] \times \{0,1\} \to [-1,1]$. Since the second argument to f is Boolean, we can write $f(v,y) = w(v)y + u(v)$. Hence,

$$\begin{aligned}\mathop{\mathbb{E}}_{J^*}[f(\mathbf{v},\mathbf{y}^*)] - \mathop{\mathbb{E}}_{J^p}[f(\mathbf{v},\mathbf{y}^p)] &= \mathop{\mathbb{E}}_{J^*}[w(\mathbf{v})\mathbf{y}^*] - \mathop{\mathbb{E}}_{J^p}[w(\mathbf{v})\mathbf{y}^p] = \mathop{\mathbb{E}}_{J^*}[w(\mathbf{v})\mathbf{y}^*] - \mathop{\mathbb{E}}_{J^p}[w(\mathbf{v})\mathbf{v}] \\ &= \mathop{\mathbb{E}}_{J^*}[w(\mathbf{v})(\mathbf{y}^* - \mathbf{v})]. \end{aligned} \tag{3.1}$$

where the first and third equalities hold because $\mathbf{v}$ is identically distributed under J^* and J^p, and the second is because $\mathbb{E}[\mathbf{y}^p|\mathbf{v}] = \mathbf{v}$. This tells us that we can limit ourselves to distinguishers of the form $f(v,y) = w(v)y$, and the distinguishing advantage can be thought of as an expectation under the single distribution J^* (Equation (3.1)). This leads to the following definition from [Gopalan et al. 2022].

Definition 3.1 *Weighted calibration error [Gopalan et al. 2022].* Let $W = \{w : [0,1] \to [-1,1]\}$ be a family of weight functions. The *W-weighted calibration error* of the predictor $p : \mathcal{X} \to [0,1]$ is defined as

$$\mathrm{CE}_W(p, \mathcal{D}^*) = \max_{w \in W} \left| \mathop{\mathbb{E}}_{\mathcal{D}^*}[w(p(\mathbf{x}))(\mathbf{y}^* - p(\mathbf{x}))] \right|.$$

The definition of weighted calibration error suggests a natural computational problem: the problem of calibration auditing for a weight family W. This is the computational problem of deciding whether $\mathrm{CE}_W(p, \mathcal{D}^*)$ is 0 or exceeds α, given access to random samples $(p(\mathbf{x}), \mathbf{y}^*)$ from $\mathcal{D}^*$. This problem turns out to be closely related to agnostic learning for the class W, as shown by [Gopalan et al. 2024].

If we instantiate weighted calibration with $W = B$ where B is the set of all bounded functions introduced in 2.2, we recover ECE. But this also illustrates why ECE is hard to compute efficiently: the set B has infinite VC dimension, hence it cannot be learnt efficiently.

Note that we could have defined the weighted calibration error CE_W as a function of J^*, the joint distribution of $(p(\mathbf{x}), \mathbf{y}^*)$, rather than the pair $(p, \mathcal{D}^*)$. We prefer mentioning p explicitly for clarity, but it is important to note that CE_W only depends on J^*. Indeed, most common measures of calibration error and loss only depend on the distribution of J^*. For instance, the cross-entropy loss and square loss only depend on how labels and predictions are jointly distributed, not on whether we are labeling images or tabular data; if we predict $p(x) = 0.7$ and the label is 1, that fixes the loss suffered at x, regardless of the features x.

3.2 Smooth calibration

Smooth calibration, introduced by [Kakade and Foster 2008] is an instantiation of weighted calibration that restricts the class of weight functions to Lipschitz continuous functions. This ensures that small perturbations of the predictor do not result in large changes in the calibration error.

Definition 3.2. Let $L = \{l : [0,1] \to [-1,1]\}$ denote the subset of 1-Lipschitz functions from B. Define the *smooth calibration error* of the predictor p under the

distribution $\mathcal{D}^*$ as $\mathrm{smCE}(p, \mathcal{D}^*) = \mathrm{CE}_L(p, \mathcal{D}^*)$.

By only allowing Lipschitz weight functions, Smooth calibration ensures that the calibration error does not change dramatically under small perturbations of the predictor.[9] Given predictors $p_1, p_2 : \mathcal{X} \to [0,1]$ and a distribution $\mathcal{D}^*$ on $\mathcal{X}$, let the expected ℓ_1 distance between them be

$$d(p_1, p_2) = \mathop{\mathbb{E}}_{\mathcal{D}^*}[|p_1(\mathbf{x}) - p_2(\mathbf{x})|].$$

Smooth calibration error is Lipschitz in this distance.

LEMMA 3.3. *For any weight family $W \subseteq L$, $\mathrm{CE}_W(p, \mathcal{D}^*)$ is 4-Lipschitz in d.*

Returning to the example above with p_0 and p_ϵ, restricting to Lipschitz distinguishers means that smooth calibration considers p_ϵ to also be well calibrated, since its smooth calibration error is $O(\epsilon)$.

An alternate view of smooth calibration is in terms of earthmover distance between J^* and J^p. Consider the ℓ_1 metric on $[0,1] \times \{0,1\}$ where $\ell_1((v,y),(v',y')) = |v - v'| + |y - y'|$. For two distributions J, J' on $[0,1] \times \{0,1\}$, we denote the earthmover distance between two distributions under the ℓ_1 metric as $\mathrm{EMD}(J, J')$. Smooth calibration captures the earth-mover distance between J^* and J^p.

LEMMA 3.4. *We have* $\mathrm{EMD}(J^*, J^p)/2 \leq \mathrm{smCE}(p, \mathcal{D}^*) \leq \mathrm{EMD}(J^*, J^p)$.

This lemma should be contrasted with Lemma 2.3, which characterizes ECE in terms of the total variation distance.

We have defined smooth calibration error in terms of the family of 1-Lipschitz distinguishers. But since an L-Lipschitz function for $L > 1$ can be made 1-Lipschitz by rescaling the range by L, the calibration error can only increase by L even if we allow L-Lipschitz distinguishers.

3.3 Other notions of weighted calibration

We have seen two notions of weighted calibration so far: ECE and smCE. Several other calibration metrics that have been considered in the literature can be naturally viewed as instances of weighed calibration. We list some of them below.

—Low-degree calibration [Gopalan et al. 2022] corresponds to the case where $W = P_d$ consists of degree d polynomials. This class is fairly Lipschitz (since polynomials have bounded derivatives. The main attraction of this notion is that it is efficient to computer, even in the multiclass setting.

—In Kernel calibration [Kumar et al. 2018; Blasiok et al. 2023a] the family of weight functions lies in a Reproducing Kernel Hilbert Space. There are many choices of kernel possible, such as the Laplace kernel, the Gaussian kernel or the polynomial kernel, each of these results in distinct calibration measures with their own properties.

[9]Note that Lemma 2.2 tells us that there exists a bounded function b_ϵ that explains the high ECE for p_ϵ, specifically, $b_\epsilon(v) = \mathrm{sign}(v - 1/2)$. This function is discontinuous near 1/2, which causes the extreme sensitivity to perturbations.

4. CALIBRATION ERROR FOR DECISION MAKING

In this section, we will explore a second approach to relaxing the definition of perfect calibration, where rather than asking J^* and J^p be identical, we require them to be close when measured under a suitable divergence. This leads to another important measure of the calibration error, the Calibration Decision Loss (CDL), introduced recently by Hu and Wu [Hu and Wu 2024]. Underlying the notion of CDL is a concrete and natural quantification of the economic value of calibration from the perspective of downstream decision making.

We define the notion of CDL in Section 4.1 and discuss its alternative formulation using Bregman divergences between J^* and J^p in Section 4.3. A key tool we use to prove this Bregman divergence formulation is a classic characterization of *proper scoring rules* [McCarthy 1956; Savage 1971; Gneiting and Raftery 2007].

4.1 Calibration Decision Loss

The definition of the Calibration Decision Loss comes naturally when we look at calibration through an economic lens, from the perspective of downstream decision makers. What does calibration mean to a person who uses the predictions (e.g. chance of rain) to make downstream decisions (e.g. take an umbrella or not)? We will show that a calibrated predictor provides a concrete *trustworthiness* guarantee to *every* payoff-maximizing downstream decision maker (Theorem 4.1). This observation gives not only a characterization of perfect calibration, but also a natural way of quantifying the calibration error of a miscalibrated predictor, using the payoff loss caused by trusting the (miscalibrated) predictor in downstream decision making. This way of quantifying the calibration error leads exactly to Calibration Decision Loss (Definition 4.2).

We start by formally defining decision tasks. A decision task $\mathcal{T}$ has two components: an action space A and a payoff function $u : A \times \{0, 1\} \to \mathbb{R}$. Given a decision task $\mathcal{T} = (A, u)$, the decision maker must pick an action $a \in A$ in order to maximize the payoff $u(a, y) \in \mathbb{R}$. Here, the payoff depends not only on the chosen action a, but also on the true outcome $y \in \{0, 1\}$ unknown to the decision maker. For example, if the outcome $y \in \{0, 1\}$ represents whether or not it will be rainy today, a natural decision task may have two actions to choose from: $A = \{\text{take umbrella}, \text{not take umbrella}\}$. Each combination (a, y) of action and outcome corresponds to a payoff value $u(a, y)$ that may depend on the susceptibility to rain and the inconvenience of carrying an umbrella.

Prediction enables decision making under uncertainty. While the decision maker is unable to observe the true outcome y before choosing the action, we assume that they are assisted by a prediction $v \in [0, 1]$. In the ideal case, the prediction correctly represents the probability distribution of the true outcome. That is, the outcome y follows the Bernoulli distribution with parameter v (denoted $\mathbf{y} \sim v$). To maximize the expected payoff, the decision maker should choose the action

$$\sigma_{\mathcal{T}}(v) \in \arg\max_{a \in A} \mathop{\mathbb{E}}_{\mathbf{y} \sim v} u(a, \mathbf{y}) \tag{4.1}$$

in response to the (correct) prediction v. We call the function $\sigma_{\mathcal{T}} : [0, 1] \to A$ the *best-response function*. Throughout the section, we assume that each decision task $\mathcal{T} = (A, u)$ is associated with a well-defined best-response function. That is, we

focus on tasks $\mathcal{T}$ where the arg max in (4.1) is always non-empty.

In reality, predictions are seldom perfectly correct. It is thus unclear whether applying the best-response function would still lead to optimal payoff. The following theorem tells us that as long as the predictions are calibrated, the best response function remains the optimal mapping from predictions to actions, allowing the decision maker to *trust the predictions as if they were correct.*

THEOREM 4.1 CALIBRATED PREDICTIONS ARE TRUSTWORTHY. *Let $\mathcal{D}$ be a joint distribution on $\mathcal{X} \times \{0,1\}$. For any perfectly calibrated predictor $p : \mathcal{X} \to [0,1]$ and any decision task $\mathcal{T} = (A, u)$, it holds that*

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x})), \mathbf{y})] = \max_{\sigma:[0,1]\to A} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma(p(\mathbf{x})), \mathbf{y})]. \tag{4.2}$$

In other words, the maximum value of the expected payoff is attained when we choose $\sigma = \sigma_{\mathcal{T}}$. Conversely, if (4.2) holds for every decision task $\mathcal{T}$, then the predictor p is perfectly calibrated.

We defer the proof of Theorem 4.1 to Section 4.2 and discuss how it suggests a new calibration measure. According to the theorem, if a predictor p is miscalibrated, then the right-hand side of (4.2) is larger than the left-hand side for some decision task $\mathcal{T}$. The difference between the two sides is exactly the payoff loss incurred by the decision maker who follows the best-response strategy $\sigma_{\mathcal{T}}$ assuming (incorrectly) that the predictions were calibrated. Thus, a natural measure of the level of miscalibration is exactly this payoff loss. For a fixed decision task $\mathcal{T}$, this payoff loss is termed the *Calibration Fixed Decision Loss (CFDL)* [Hu and Wu 2024]. Taking the worst-case payoff loss over all decision tasks $\mathcal{T} = (A, u)$ with bounded payoff functions $u : A \to [0,1]$, we get the Calibration Decision Loss (CDL).

Definition 4.2 *Calibration Decision Loss (CDL) [Hu and Wu 2024].* Let $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \{0,1\}$. Given a predictor $p : \mathcal{X} \to [0,1]$, we define its *Calibration Fixed Decision Loss (CFDL)* with respect to a (fixed) decision task $\mathcal{T} = (A, u)$ as

$$\mathrm{CFDL}_{\mathcal{T}}(p, \mathcal{D}) := \max_{\sigma:[0,1]\to A} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma(p(\mathbf{x})), \mathbf{y})] - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x})), \mathbf{y})].$$

We define the *Calibration Decision Loss (CDL)* of the predictor p as the supremum of the CFDL over all decision tasks (A, u) where the payoff function $u : A \to [0,1]$ has its range bounded in $[0,1]$:

$$\mathrm{CDL}(p, \mathcal{D}) := \sup_{\mathcal{T}=(A,u), u:A\to[0,1]} \mathrm{CFDL}_{\mathcal{T}}(p, \mathcal{D}).$$

As we will see when we prove Theorem 4.1 in Section 4.2, the CDL is zero if and only if the predictor p is perfectly calibrated. If a predictor is not perfectly calibrated but has a small CDL, any decision maker can still trust the predictor as if it were calibrated without losing too much expected payoff. This holds because the CDL is the supremum of the CFDL over *all* payoff-bounded decision tasks.

We note that in the definition of CDL, decision tasks are restricted to have a bounded payoff function $u : A \to [0,1]$. This restriction is only for the purpose of normalization: multiplying the payoff function by any positive constant changes the corresponding CFDL by the same constant factor, whereas adding a constant

to the payoff function does not change the CFDL. There is no further restriction on the decision tasks beyond bounded payoff functions. In particular, the action set A can have arbitrary (even infinite) size. A small CDL implies that trusting the predictions will incur small payoff loss for *all* such decision tasks.

A natural question is how the CDL is related to other measures of the calibration error. We will prove that the CDL is quadratically related to the ECE:

THEOREM 4.3 [KLEINBERG ET AL. 2023; HU AND WU 2024]. *Let $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \{0,1\}$. For any predictor $p : \mathcal{X} \to [0,1]$,*

$$\mathrm{ECE}(p,\mathcal{D})^2 \le \mathrm{ECE}_2(p,\mathcal{D})^2 \le \mathrm{CDL}(p,\mathcal{D}) \le 2\,\mathrm{ECE}(p,\mathcal{D}) \le 2\,\mathrm{ECE}_2(p,\mathcal{D}). \quad (4.3)$$

Moreover, the quadratic relationship between CDL and ECE shown in Theorem 4.3 is tight (up to lower order terms): for any $\varepsilon \in (0, 1/10)$, there exist two pairs $(p_1,\mathcal{D}_1),(p_2,\mathcal{D}_2)$ such that

$$\mathrm{ECE}_2(p_1,\mathcal{D}_1) = \varepsilon, \qquad \mathrm{CDL}(p_1,\mathcal{D}_1) = 2\varepsilon;$$
$$\mathrm{ECE}(p_2,\mathcal{D}_2) = \varepsilon, \qquad \mathrm{CDL}(p_2,\mathcal{D}_2) \le \varepsilon^2 + O(\varepsilon^3).$$

We defer the proof of Theorem 4.3 to Section A.7. Here we briefly describe the two tight examples. The first example $(p_1,\mathcal{D}_1)$ is very simple. For $(\mathbf{x},\mathbf{y}) \sim \mathcal{D}_1$, we draw $\mathbf{y} \in \{0,1\}$ from the Bernoulli distribution with parameter $1/2+\varepsilon$, independent of $\mathbf{x}$. The predictor p_1 is the constant predictor $p_1(x) = 1/2$. In the second example, we draw $\mathbf{x}$ uniformly at random from the interval $[\varepsilon,1]$ and then draw $\mathbf{y} \in \{0,1\}$ from the Bernoulli distribution with parameter $\mathbf{x}-\varepsilon$. The predictor p_2 is the identity function $p_2(x) = x$ for $x \in [\varepsilon,1]$. We will prove the correctness of the examples in Section A.8.

The second example, $(p_2,\mathcal{D}_2)$, demonstrating that the CDL can be significantly smaller than the ECE, is quite instructive. It opens up the possibility that the CDL can be minimized at a faster rate than what is possible for ECE in the online setting. Indeed, the main technical result of [Hu and Wu 2024] gives an efficient online CDL minimization algorithm achieving rate $O(\sqrt{T}\log T)$, overcoming the information-theoretic lower bound $\Omega(T^{0.54389})$ for ECE [Qiao and Valiant 2021; Dagan et al. 2025] (see Section 5 for more discussions).

To conclude this subsection, CDL measures the calibration error using the payoff loss of downstream decision makers caused by mis-calibration. In addition to introducing CDL as a meaningful decision-theoretic measure of calibration, the work of [Hu and Wu 2024] also shows that CDL allows a significantly better rate than what is possible for ECE in online calibration, which we discuss in Section 5.

In Section 4.2 we give a simpler yet equivalent definition of the CFDL in (4.5), which leads to an interpretation of CDL through the lens of indistinguishability.

4.2 Characterization of the Maximum Expected Payoff

In this section we prove Theorem 4.1. We start by giving a characterization of the maximum expected payoff on the right-hand side of (4.2) for a general predictor p that may or may not be calibrated, which simplifies the definition of CFDL and will be useful in the proof.

Recall the definition of the recalibration $\widehat{p}$ of p (Definition 6.4): $\widehat{p}$ is obtained by replacing each prediction value $v = p(x)$ with the actual conditional expectation

$\mathbb{E}[\mathbf{y}|p(\mathbf{x}) = v]$. Clearly, $\widehat{p}$ is perfectly calibrated. If p is perfectly calibrated, then $\widehat{p} = p$. We have the following characterization of the maximum expected payoff achievable by post-processing p (see Section A.3 for proof):

LEMMA 4.4. *Let $\mathcal{D}$ be a joint distribution on $\mathcal{X} \times \{0,1\}$. For any predictor $p : \mathcal{X} \to [0,1]$ and any decision task $\mathcal{T} = (A, u)$, it holds that*

$$\max_{\sigma:[0,1]\to A} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma(p(\mathbf{x})), \mathbf{y})] = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x})), \mathbf{y})], \tag{4.4}$$

where $\widehat{p}$ is the recalibration of p.

We can now rewrite the definition of CFDL (Definition 4.2) based on Lemma 4.4:

$$\mathrm{CFDL}_{\mathcal{T}}(p, \mathcal{D}) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x})), \mathbf{y})] - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x})), \mathbf{y})]. \tag{4.5}$$

This expression allows us to easily calculate the CFDL for specific decision tasks. For example, consider the task $\mathcal{T}_2 = (A, u)$ where the action space A is the unit interval $A = [0,1]$, and the payoff function is quadratic:

$$u(a, y) = 1 - (a - y)^2 \in [0,1], \quad \text{for } a \in [0,1] \text{ and } y \in \{0,1\}.$$

The corresponding best-response function is the identity: $\sigma_{\mathcal{T}}(v) = v$. Plugging it in (4.5), we obtain an equality between the CFDL and the square of ECE_2:

$$\begin{aligned}
\mathrm{CFDL}_{\mathcal{T}_2}(p, \mathcal{D}) &= \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(p(\mathbf{x}) - \mathbf{y})^2 - (\widehat{p}(\mathbf{x}) - \mathbf{y})^2] \\
&= \mathbb{E}[p(\mathbf{x})^2 - \widehat{p}(\mathbf{x})^2 + 2\mathbf{y}(\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \\
&= \mathbb{E}[p(\mathbf{x})^2 - \widehat{p}(\mathbf{x})^2 + 2\widehat{p}(\mathbf{x})(\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \quad (\mathbb{E}[\mathbf{y}|\widehat{p}(\mathbf{x}), p(\mathbf{x})] = \widehat{p}(\mathbf{x})) \\
&= \mathbb{E}[(p(\mathbf{x}) - \widehat{p}(\mathbf{x}))^2] = \mathrm{ECE}_2(p, \mathcal{D})^2.
\end{aligned} \tag{4.6}$$

We are now ready to prove Theorem 4.1.

PROOF OF THEOREM 4.1. If p is perfectly calibrated, then $p = \widehat{p}$, and (4.2) follows immediately from Lemma 4.4. For the reverse direction, if (4.2) holds for any decision task, then in particular, it holds for the task $\mathcal{T}_2$ above, implying $\mathrm{CFDL}_{\mathcal{T}_2}(p, \mathcal{D}) = 0$. By (4.6), we have $\mathrm{ECE}_2(p, \mathcal{D}) = 0$, so p is perfectly calibrated, as desired. Since the quadratic payoff function of $\mathcal{T}_2$ has a bounded range $[0,1]$, this proof also implies that the CDL of a predictor is zero if and only if the predictor is perfectly calibrated. □

4.3 The Bregman Divergence View of CDL

We show that the CFDL of a predictor p w.r.t. any decision task $\mathcal{T}$ can be expressed as a Bregman divergence $D_\varphi(J^*\|J^p)$ between the two joint distributions J^* and J^p (Theorem 4.9). Our proof uses a classic characterization of *proper scoring rules* [McCarthy 1956; Savage 1971; Gneiting and Raftery 2007].

We start with the definition of Bregman divergence.

Definition 4.5 *Bregman Divergence.* Let $\varphi : [0,1] \to \mathbb{R}$ be a convex function and let $\nabla\varphi : [0,1] \to \mathbb{R}$ be its subgradient. For any pair of values $\mu^*, \mu \in [0,1]$, their *Bregman divergence* w.r.t. φ is

$$D_\varphi(\mu^*\|\mu) := \varphi(\mu^*) - \varphi(\mu) - \nabla\varphi(\mu) \cdot (\mu^* - \mu).$$

Since $\nabla\varphi(\mu)$ is a subgradient of φ at μ, the Bregman divergence is always nonnegative. When $\mu = \mu^*$, the Bregman divergence becomes zero.

We will interpret the values $\mu^*, \mu \in [0,1]$ in the definition above as the parameters of two Bernoulli distributions. For example, if we choose $\varphi(\mu)$ to be the negative Shannon entropy of the Bernoulli distribution with parameter μ:

$$\varphi(\mu) = \mu \ln \mu - (1-\mu)\ln(1-\mu),$$

then the Bregman divergence becomes the KL divergence between the two Bernoulli distributions parameterized by μ^* and μ:

$$D_\varphi(\mu^* \| \mu) = \mu^* \ln \frac{\mu^*}{\mu} + (1-\mu^*) \ln \frac{1-\mu^*}{1-\mu}.$$

The following key theorem makes the connection between Bregman divergences and decision tasks.

Theorem 4.6. *For any decision task $\mathcal{T} = (A, u)$, there exists a convex function $\varphi : [0,1] \to \mathbb{R}$ with subgradient $\nabla\varphi : [0,1] \to \mathbb{R}$ such that*

$$u(\sigma_\mathcal{T}(v), y) = \varphi(v) + \nabla\varphi(v) \cdot (y - v) \quad \textit{for every } v \in [0,1] \textit{ and } y \in \{0,1\}.$$

To prove the theorem, one should first observe that the function $u(\sigma_\mathcal{T}(v), y)$ is a *proper scoring rule*. That is, for any $v, v' \in [0,1]$, we have

$$\mathop{\mathbb{E}}_{\mathbf{y}\sim v} u(\sigma_\mathcal{T}(v), \mathbf{y}) \geq \mathop{\mathbb{E}}_{\mathbf{y}\sim v} u(\sigma_\mathcal{T}(v'), \mathbf{y}),$$

which follows from the definition (4.1) of the best-response function $\sigma_\mathcal{T}$. The theorem then follows from a standard characterization of proper scoring rules [McCarthy 1956; Savage 1971; Gneiting and Raftery 2007].

We can now write the expected payoff achieved by a predictor p using the Bregman divergence between p and its recalibration $\widehat{p}$ (see Section A.4 for proof):

Lemma 4.7. *Fix a joint distribution $\mathcal{D}$ of $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \{0,1\}$. Let $p : \mathcal{X} \to [0,1]$ be a predictor and $\widehat{p}$ be its recalibration (Definition 6.4). Then for any decision task $\mathcal{T} = (A, u)$ and the corresponding convex function φ from Theorem 4.6,*

$$\mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_\mathcal{T}(p(\mathbf{x})), \mathbf{y})] = \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x}) \| p(\mathbf{x}))], \tag{4.7}$$

$$\mathrm{CFDL}_\mathcal{T}(p, \mathcal{D}) = \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x}) \| p(\mathbf{x}))]. \tag{4.8}$$

We now generalize the definition of Bregman divergence to joint distributions, such as J^* and J^p, over the domain $[0,1] \times \{0,1\}$.

Definition 4.8 *Induced Bregman Divergence between Joint Distributions.* Let $\varphi : [0,1] \to \mathbb{R}$ be a convex function and let $\nabla\varphi : [0,1] \to \mathbb{R}$ be its subgradient. For any joint distribution J of $(\mathbf{v}, \mathbf{y}) \in [0,1] \times \{0,1\}$, we use $\mu_J(\mathbf{v}) = \mathbb{E}_J[\mathbf{y}|\mathbf{v}] \in [0,1]$ to denote the parameter of the Bernoulli distribution of $\mathbf{y}$ conditioned on $\mathbf{v}$. Let J_1, J_2 be a pair of joint distributions of $(\mathbf{v}, \mathbf{y}) \in [0,1] \times \{0,1\}$ that share the same marginal distribution of $\mathbf{v}$ and denote this marginal distribution by M. We define

the *Bregman divergence* between J_1 and J_2 *induced by* φ as[10]

$$D_\varphi(J_1 \| J_2) := \mathop{\mathbb{E}}_{\mathbf{v} \sim M}[D_\varphi(\mu_{J_1}(\mathbf{v}) \| \mu_{J_2}(\mathbf{v}))].$$

Combining Lemma 4.7 and Definition 4.8, we have a Bregman divergence characterization of the CFDL for any decision task $\mathcal{T}$ (see Section A.5 for proof).

THEOREM 4.9 BREGMAN DIVERGENCE VIEW OF CFDL. *Let $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \{0,1\}$, and let $p : \mathcal{X} \to [0,1]$ be a predictor. As before, given $(\mathbf{x}, \mathbf{y}^*) \sim \mathcal{D}$, we draw $\mathbf{y}^p$ from the Bernoulli distribution with parameter $p(\mathbf{x})$, and use J^*, J^p to denote the distributions of $(p(\mathbf{x}), \mathbf{y}^*)$ and $(p(\mathbf{x}), \mathbf{y}^p)$, respectively. Then for any decision task $\mathcal{T} = (A, u)$ and the corresponding convex function φ from Theorem 4.6,* $\mathrm{CFDL}_\mathcal{T}(p, \mathcal{D}) = D_\varphi(J^* \| J^p)$.

4.4 Further Work

As we discuss in this section, the defining property of CDL is that it provides a meaningful guarantee on the swap regret incurred by downstream decision makers who trust the predictions. However, CDL is undesirable in other aspects: like ECE, it is discontinuous and requires high sample complexity to estimate. Recent work of [Rossellini et al. 2025] introduces the notion of *cutoff calibration error* to address the sample complexity issue while maintaining a restricted form of the decision-theoretic guarantee of CDL (e.g. they consider the regret relative to *monotone* post-processings of the predictions). This notion of cutoff calibration is essentially identical to the notion of *proper calibration* from [Okoroafor et al. 2025], who give an algorithm achieving $\widetilde{O}(\sqrt{T})$ error rate for proper calibration in the online setting (see Section 5 for the setting). The works of [Blasiok et al. 2023b; Blasiok and Nakkiran 2024; Hartline et al. 2025] show that low smooth calibration error also gives certain decision-theoretic guarantees. In particular, these works show that it implies low regret for certain forms of Lipschitz post-processings or for decision makers who make randomized responses (e.g. by adding noise to the predictions), though this implication often comes with a quantitative loss (e.g. smooth calibration error being at most ε only implies an $O(\sqrt{\varepsilon})$ regret).

5. ONLINE CALIBRATION

We have discussed a variety of ways to quantify the calibration error of a given predictor. In this section, we turn to the algorithmic question of *constructing* a predictor with low calibration error. This question, when naively formulated, admits a trivial and unenlightening solution: one can simply construct a constant predictor that assigns (an approximation of) the overall average $\mathbb{E}[y]$ to every individual x. This is a well-calibrated predictor according to every calibration measure we have discussed. Thus, for the algorithmic question to be insightful, it is essential to formulate it in such a way that reaches beyond the trivial solution. The seminal work of Foster and Vohra [Foster and Vohra 1998] introduced one such interesting question that turned into an active area of research with exciting recent progress:

[10] One can also view $D_\varphi(J_1 \| J_2)$ as the Bregman divergence corresponding to the negative entropy $\Phi(J)$ of any joint distribution J of $(\mathbf{v}, \mathbf{y}) \in [0,1] \times \{0,1\}$ defined by $\Phi(J) := \mathbb{E}_{(\mathbf{v},\mathbf{y}) \sim J}[\varphi(\mu_J(\mathbf{v}))]$.

calibration in *online* prediction. We will first describe the problem setting and then briefly survey some key results in the literature.

The online prediction problem has T rounds indexed by $t \in [T]$. In round t, our algorithm makes a prediction $p_t \in [0,1]$, and nature reveals an outcome $y_t \in \{0,1\}$. For example, we can interpret the problem as predicting the chance of rain each day for T days, where p_t is the prediction we make on day t, and $y_t = 1$ if day t is rainy. Since the rounds are ordered chronologically, we allow our algorithm to choose p_t as a function of the history $H_{t-1} = (p_1, \dots, p_{t-1}, y_1, \dots, y_{t-1})$, and similarly, y_t can depend on the history H_{t-1} as well.

To evaluate the calibration error of the prediction sequence $p_{1,\dots,T} := (p_1, \dots, p_T)$ w.r.t. the outcome sequence $y_{1,\dots,T} := (y_1, \dots, y_T)$, we consider the predictor $p : \{1, \dots, T\} \to [0,1]$ that assigns prediction $p(t) := p_t$ to each time step $t = 1, \dots, T$. Viewing each time step as an individual, we let $\mathcal{D}$ be the uniform distribution over the individual-outcome pairs (t, y_t) for $t = 1, \dots, T$. By slight abuse of notation, we can transform any calibration measure CAL for $(p, \mathcal{D})$ into a calibration measure CAL for $(p_{1,\dots,T}, y_{1,\dots,T})$ as follows:

$$\mathrm{CAL}(p_{1,\dots,T}, y_{1,\dots,T}) := T\,\mathrm{CAL}(p, \mathcal{D}).$$

Once a calibration measure CAL is chosen, our goal is to design a prediction algorithm that guarantees a small (e.g. sub-linear, i.e., $o(T)$) calibration error according to CAL, regardless of how the outcomes y_t are generated. We wish to design a prediction algorithm P that specifies how p_t should be chosen as a function of the history H_{t-1} for every round t. We want the calibration error to be small regardless of nature's strategy Y, which specifies how y_t should be chosen as a function of H_{t-1} for every round t. That is, we want to solve the following optimization problem:

$$\operatorname*{minimize}_{P} \max_{Y} \mathrm{CAL}(p_{1,\dots,T}, y_{1,\dots,T}), \text{ where } p_{1,\dots,T}, y_{1,\dots,T} \text{ is generated by } P \text{ and } Y.$$

For some calibration measures (e.g. ECE and CDL), it is necessary to use randomized prediction algorithms to achieve sub-linear rates. Such an algorithm constructs a distribution $\mathcal{P}$ over prediction strategies P to solve the following problem:

$$\operatorname*{minimize}_{\mathcal{P}} \max_{Y} \mathop{\mathbb{E}}_{P \sim \mathcal{P}}[\mathrm{CAL}(p_{1,\dots,T}, y_{1,\dots,T})].$$

Here is why randomized predictions are necessary for achieving sub-linear rates for ECE or CDL. For every deterministic prediction algorithm P, nature can infer the prediction p_t based on the history H_{t-1}, and can then choose $y_t = 1$ if and only if $p_t < 1/2$, incurring an $\Omega(T)$ rate for ECE and CDL.

In Table I, we summarize the current best upper and lower bounds on the optimal online calibration rates for a few calibration error measures we discussed earlier, which is an active topic for recent research. Notably, the only calibration measure in this table that does not allow an $\widetilde{O}(\sqrt{T})$ rate is ECE.

There are substantial gaps between the best upper and lower bounds for many calibration measures in this table, making it a natural question to close or reduce these gaps. Very recently, the works of [Peng 2025] and [Fishelson et al. 2025] have achieved significant progress on online calibration algorithms in the *multi-class* setting, opening up another exciting area for future research.

Calibration Error	Rate Upper Bound	Rate Lower Bound
Expected Calibration Error (ECE)	$O(T^{2/3})$ [Foster and Vohra 1998]	$\Omega(T^{1/2})$ [Folklore]
		$\Omega(T^{0.528})$ [Qiao and Valiant 2021]
	$O(T^{2/3-\varepsilon})$ [Dagan et al. 2025]	$\Omega(T^{0.54389})$ [Dagan et al. 2025]
Distance to Calibration [Blasiok et al. 2023a]	$O(T^{1/2})$ [Qiao and Zheng 2024] [Arunachaleswaran et al. 2025]	$\Omega(T^{1/3})$ [Qiao and Zheng 2024]
Smooth Calibration Error [Kakade and Foster 2008]	$O(T^{1/2})$ [Qiao and Zheng 2024] [Arunachaleswaran et al. 2025]	$\Omega(T^{1/3})$ [Qiao and Zheng 2024]
Calibration Decision Loss (CDL) [Hu and Wu 2024]	$O(T^{1/2}\log T)$ [Hu and Wu 2024]	$\Omega(T^{1/2})$ [Hu and Wu 2024]

Table I. Upper and lower bounds on the optimal rates for online calibration

6. THE DISTANCE TO CALIBRATION

At this point, we seem to have a Cambrian explosion of approximate calibration measures, each of which has their own desirable properties, and will give different calibration errors for a predictor. How should we compare these different measures, and decide which to use? Is there any notion of ground truth, that would guide us in this choice? In this section, we present one possible answer to this question via the notion of the distance to calibration [Blasiok et al. 2023a]. We show that the smooth calibration error gives us the best approximation to this ground-truth measure in an information-theoretic sense.

Recall that we defined $\mathcal{D}^*$ to be the joint distribution of $\mathbf{x}, \mathbf{y}^*$, whereas J^* denotes the joint distribution $(p(\mathbf{x}), \mathbf{y}^*)$.

Definition 6.1 *Distance to calibration [Blasiok et al. 2023a].* Given a distribution $\mathcal{D}^*$, define $\mathrm{Cal}(\mathcal{D}^*)$ to be the set of predictors $q : \mathcal{X} \to [0,1]$ such that q is perfectly calibrated under $\mathcal{D}^*$. Define the *true distance to calibration* of the predictor p as

$$\mathrm{dCE}(p, \mathcal{D}^*) = \min_{q \in \mathrm{Cal}(D^*)} d(p,q).$$

This definition formalizes the intuition that a predictor which can be made perfectly calibrated by a small change to its predictions is close to being calibrated. A desirable property that follows immediately from this definition is that the distance to calibration is continuous (unlike ECE). In fact, dCE is Lipschitz continuous: if we chance our predictor p to a different predictor p' that is ε-close to p ($|d(p,p') \leq \varepsilon|$), the distance to calibration can only change by at most ε ($|\mathrm{dCE}(p,\mathcal{D}^*) - \mathrm{dCE}(p',\mathcal{D}^*)| \leq \varepsilon$). This continuity property can be easily proved using the triangle inequality for the metric d.

Despite its intuitiveness and continuity, dCE differs from the other notions of calibration we have seen so far in a crucial way: it depends on the feature space $\mathcal{X}$ (at least, syntactically). This dependence comes about because both the set $\mathrm{Cal}(\mathcal{D}^*)$ of perfectly calibrated predictors and the distance metric d depend on $\mathcal{X}$. The definition of dCE does not give any hints about how one might go about

computing or approximating it.

It is natural to ask to what extent dCE actually depends on the space $\mathcal{X}$, and if it can be approximated by a calibration measure which is independent of $\mathcal{X}$. This leads us to two new definitions.

Definition 6.2 *[Blasiok et al. 2023a].* The *upper distance to calibration* $\overline{\mathrm{dCE}}(J^*)$ is the maximum of $\mathrm{dCE}(p', \mathcal{D}')$ over all spaces $\mathcal{X}'$, distributions $\mathcal{D}'$ on $\mathcal{X}' \times \{0,1\}$ and predictors $p' : \mathcal{X}' \to [0,1]$ such that the distribution $J' = (p'(\mathbf{x}'), \mathbf{y}')$ is identical to the distribution $J^* = (\mathbf{p}(\mathbf{x}), \mathbf{y}^*)$. The *lower distance to calibration* $\underline{\mathrm{dCE}}$ is defined analogously, replacing the maximum by minimum.

By their definition, both $\underline{\mathrm{dCE}}$ and $\overline{\mathrm{dCE}}$ achieve the goal of only depending on J^* and not $\mathcal{D}^*$. It also follows that

$$\underline{\mathrm{dCE}}(J^*) \le \mathrm{dCE}(p, \mathcal{D}^*) \le \overline{\mathrm{dCE}}(J^*).$$

This leads to two questions:

(1) The definitions of $\underline{\mathrm{dCE}}$ and $\overline{\mathrm{dCE}}$ seem rather cumbersome at first, since they involve optimizing over a possibly infinite collection of feature spaces and predictors. Are there more tractable characterizations of these notions, ideally ones that will let us estimate them efficiently?

(2) How far apart are $\underline{\mathrm{dCE}}$ and $\overline{\mathrm{dCE}}$? An ideal situation would be that they are always equal, or at most a constant factor apart. If so, either of them could serve as a good approximation for dCE, assuming we find efficient ways to compute them.

In the following subsection, we will show that the largest gap between the upper and lower distances is quadratic ($\overline{\mathrm{dCE}}(J^*) \le 4\sqrt{\underline{\mathrm{dCE}}(J^*)}$), and that the smooth calibration error gives a constant-factor approximation to the lower distance to calibration. Together, these results let us efficiently approximate the distance to calibration using smooth calibration error, as in the work of [Hu et al. 2024].

6.1 Characterizing and Relating the Upper and Lower Distances to Calibration

In this subsection, we answer the two questions above. Specifically, we give simple characterizations for the upper and lower distances in Theorems 6.6 and 6.7. We show that the two distances are at most quadratically apart in Theorem 6.9.

We first give a simpler characterization of the upper distance. We begin with some definitions needed to state the characterization.

Definition 6.3 *Calibrated post-processing.* Define the set $K(J^*)$ to be the set of post-processing functions that, when applied to p, give a perfectly calibrated predictor. Formally, $K(J^*) = \{\kappa : [0,1] \to [0,1] \; s.t. \; (\kappa(p(\mathbf{x})), y^*)$ is perfectly calibrated.$\}$

We observe that the set $K(J^*)$ is non-empty, since the constant predictor which predicts $\mathbb{E}[y^*]$ is calibrated, and this corresponds to the constant function $\kappa^{\mathsf{av}}(v) = \mathbb{E}[\mathbf{y}^*]$ for all v. A more interesting post-processing is $\kappa^{\mathsf{recal}}(v) = \mathbb{E}[\mathbf{y}^* | v]$, and we call the post-processed predictor $\widehat{p}(\mathbf{x}) := \kappa^{\mathsf{recal}}(p(\mathbf{x}))$ the *recalibration* of p: this predictor keeps the same level sets as p, and changes the predictions to be calibrated.

Definition 6.4 *Recalibration.* Fix a distribution $\mathcal{D}$ of $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \{0,1\}$. We define the *recalibration* of a predictor $p : \mathcal{X} \to [0,1]$ to be another predictor, denoted by $\widehat{p} : \mathcal{X} \to [0,1]$, where $\widehat{p}(\mathbf{x}) := \mathbb{E}_{\mathcal{D}}[\mathbf{y}|p(\mathbf{x})]$.

LEMMA 6.5. *It holds that* $\mathrm{ECE}(p, \mathcal{D}^*) = d(p, \widehat{p}) = d(p, \kappa^{\mathsf{recal}} \circ p)$*, where* $\circ$ *denotes function composition.*

In general, the set $K(J^*)$ could be much richer and possibly induce *closer* calibrated predictors. In particular, there often exist post-processings $\kappa \in K(J^*)$ such that $d(p, \kappa \circ p)$ is much smaller than $d(p, \kappa^{\mathsf{recal}} \circ p) = \mathrm{ECE}(p, \mathcal{D}^*)$. For the two point distribution $\mathcal{D}_2$ considered before, we have seen that $\mathrm{ECE}(p, \mathcal{D}_2) = 1/2 - \epsilon$ whereas it follows that $\kappa^{av} = 1/2$ and $d(p, 1/2) = \epsilon$.

[Blasiok et al. 2023a] give the following characterization of the upper distance.

THEOREM 6.6. *[Blasiok et al. 2023a] We have*

$$\overline{\mathrm{dCE}}(J^*) = \min_{\kappa \in K(J^*)} d(p, \kappa \circ p) = \min_{\kappa \in K(J^*)} \mathop{\mathbb{E}}_{\mathbf{x}} |\kappa(p(\mathbf{x})) - p(\mathbf{x})|.$$

This theorem tells us that the upper distance of a given predictor p is exactly its distance to the closest perfectly calibrated predictor that can be obtained by applying a post-processing κ to p.

Let us sketch the proof idea. $K(J^*)$ is the set of relabelings of the level sets of p which result in a calibrated predictor. For any space X', distribution D' and predictor p' where $J' = J^*$, applying the post-processing function $\kappa \in J^*$ results in a perfectly calibrated predictor $\kappa(p')$ on $\mathcal{X}'$. Hence the distance from such predictors is always an upper bound on $\overline{\mathrm{dCE}}$. For the space X'' where each level set is a single point, these are the only calibrated predictors, so the bound is tight.

We now turn to the lower distance. The good news is that the characterization is in terms of a calibration measure that we have encountered previously: the smooth calibration error $\mathrm{smCE}(p, \mathcal{D}^*)$. The proof however is more involved, we refer the reader to [Blasiok et al. 2023a; Blasiok and Nakkiran 2024].

THEOREM 6.7 [BLASIOK ET AL. 2023A]. *We have*

$$\mathrm{smCE}(p, \mathcal{D}^*)/2 \leq \underline{\mathrm{dCE}}(J^*) \leq 2\mathrm{smCE}(p, \mathcal{D}^*)$$

This theorem lets us efficiently approximate the lower distance to calibration, up to a constant factor, by computing the smooth calibration error. An efficient algorithm for computing the smooth calibration error is given by [Hu et al. 2024].

We now address the question of how close the upper and lower distances are. Assume that all we know about the predictor p and distribution $\mathcal{D}^* = (\mathbf{x}, \mathbf{y}^*)$ is the distribution $J^* = (p(\mathbf{x}), \mathbf{y}^*)$. Does this specify $\mathrm{dCE}(p, D^*)$ completely? Or is there still some uncertainty about how far the closest calibrated predictor is, depending on the space $\mathcal{X}$? The answer (perhaps surprisingly) is that there is quadratic uncertainty in the distance, given J^*.

COROLLARY 6.8. *No calibration measure based on* J^* *can distinguish between the cases where* $\mathrm{dCE}(p, \mathcal{D}^*) \geq \eta$ *and* $\mathrm{dCE}(p, \mathcal{D}^*) \leq 2\eta^2$.

We present an example illustrating Corollary 6.8 in Appendix B. Specifically, we construct pairs of predictors and distributions $(p_1, \mathcal{D}_1^*)$ and $(p_2, \mathcal{D}_2^*)$ so that J^* is

identical in both cases, but dCE differs by a quadratic factor. It turns out that this quadratic separation is in fact the worst possible.

THEOREM 6.9 [BLASIOK ET AL. 2023A]. *We have* $\overline{\mathrm{dCE}}(J^*) \leq 4\sqrt{\underline{\mathrm{dCE}}(J^*)}$.

We discuss the proof of this theorem in Appendix C following the original approach of [Blasiok et al. 2023a] via the notion of *interval calibration error*.

Conclusion.

The classic notion of calibration needs to be rethought in order to satisfy requirements like robustness and computational efficiency, motivated by applications to machine learning and decision making. This leads to a rich set of new questions, in terms of what are desirable properties for approximate calibration notions to have and new algorithmic challenges that arise from trying to achieve these properties. This is a broad and active area of research that spans machine learning, decision making and computational complexity. There are several questions that still remain, such as efficient and meaningful notions of calibration for the multiclass setting [Gopalan et al. 2024] and the generative setting [Kalai and Vempala 2024]. We hope to have given the reader a feel for this in the survey, by highlighting the motivating questions, the definitional challenges and the algorithmic issues.

REFERENCES

ARUNACHALESWARAN, E. R., COLLINA, N., ROTH, A., AND SHI, M. 2025. An elementary predictor obtaining $2\sqrt{T}+1$ distance to calibration. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2025, New Orleans, LA, USA, January 12-15, 2025*, Y. Azar and D. Panigrahi, Eds. SIAM, 1366–1370.

BLASIOK, J., GOPALAN, P., HU, L., AND NAKKIRAN, P. 2023a. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023*. ACM, 1727–1740.

BLASIOK, J., GOPALAN, P., HU, L., AND NAKKIRAN, P. 2023b. When does optimizing a proper loss yield calibration? 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., 72071–72095.

BLASIOK, J. AND NAKKIRAN, P. 2024. Smooth ECE: principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

CASACUBERTA, S., DWORK, C., AND VADHAN, S. 2024. Complexity-theoretic implications of multicalibration. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Association for Computing Machinery, New York, NY, USA, 1071–1082.

DAGAN, Y., DASKALAKIS, C., FISHELSON, M., GOLOWICH, N., KLEINBERG, R., AND OKOROAFOR, P. 2025. Breaking the t^(2/3) barrier for sequential calibration. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23-27, 2025*, M. Koucký and N. Bansal, Eds. ACM, 2007–2018.

DWORK, C., KIM, M. P., REINGOLD, O., ROTHBLUM, G. N., AND YONA, G. 2021. Outcome indistinguishability. In *ACM Symposium on Theory of Computing (STOC'21)*.

DWORK, C., LEE, D., LIN, H., AND TANKALA, P. 2023. From pseudorandomness to multi-group fairness and back. In *Proceedings of Thirty Sixth Conference on Learning Theory*, G. Neu and L. Rosasco, Eds. Proceedings of Machine Learning Research, vol. 195. PMLR, 3566–3614.

FISHELSON, M., GOLOWICH, N., MOHRI, M., AND SCHNEIDER, J. 2025. High-dimensional calibration from swap regret. *arXiv preprint arXiv:2505.21460*.

FOSTER, D. P. AND VOHRA, R. V. 1998. Asymptotic calibration. *Biometrika 85,* 2, 379–390.

GNEITING, T. AND RAFTERY, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102,* 477, 359–378.

Gopalan, P., Hu, L., and Rothblum, G. N. 2024. On computationally efficient multi-class calibration. In *The Thirty Seventh Annual Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 247. PMLR, 1983–2026.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. 2022. Omnipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*.

Gopalan, P., Kim, M. P., Singhal, M., and Zhao, S. 2022. Low-degree multicalibration. In *Conference on Learning Theory, 2-5 July 2022, London, UK*. Proceedings of Machine Learning Research, vol. 178. PMLR, 3193–3234.

Hartline, J., Wu, Y., and Yang, Y. 2025. Smooth Calibration and Decision Making. In *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, M. Bun, Ed. Leibniz International Proceedings in Informatics (LIPIcs), vol. 329. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 16:1–16:26.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML*.

Hu, L., Jambulapati, A., Tian, K., and Yang, C. 2024. Testing calibration in nearly-linear time. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hu, L. and Wu, Y. 2024. Predict to minimize swap regret for all payoff-bounded tasks. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*. 244–263.

Kakade, S. and Foster, D. 2008. Deterministic calibration and Nash equilibrium. *Journal of Computer and System Sciences 74(1)*, 115–130.

Kalai, A. T. and Vempala, S. S. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Association for Computing Machinery, New York, NY, USA, 160–171.

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., and Reingold, O. 2022. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences 119,* 4.

Kleinberg, B., Leme, R. P., Schneider, J., and Teng, Y. 2023. U-calibration: Forecasting for an unknown agent. In *Proceedings of Thirty Sixth Conference on Learning Theory*, G. Neu and L. Rosasco, Eds. Proceedings of Machine Learning Research, vol. 195. PMLR, 5143–5145.

Kumar, A., Sarawagi, S., and Jain, U. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research, vol. 80. PMLR, 2805–2814.

Li, Y., Hartline, J. D., Shan, L., and Wu, Y. 2022. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Association for Computing Machinery, New York, NY, USA, 988–989.

McCarthy, J. 1956. Measures of the value of information. *Proceedings of the National Academy of Sciences 42,* 9, 654–655.

Okoroafor, P., Kleinberg, R., and Kim, M. P. 2025. Near-optimal algorithms for omniprediction. *arXiv preprint arXiv:2501.17205*.

Peng, B. 2025. High dimensional online calibration in polynomial time. *arXiv preprint arXiv:2504.09096*.

Qiao, M. and Valiant, G. 2021. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2021. Association for Computing Machinery, New York, NY, USA, 456–466.

Qiao, M. and Zheng, L. 2024. On the distance from calibration in sequential prediction. In *Proceedings of Thirty Seventh Conference on Learning Theory*, S. Agrawal and A. Roth, Eds. Proceedings of Machine Learning Research, vol. 247. PMLR, 4307–4357.

Rossellini, R., Soloff, J. A., Barber, R. F., Ren, Z., and Willett, R. 2025. Can a calibration metric be both testable and actionable? *arXiv preprint arXiv:2502.19851*.

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66,* 336, 783–801.

A. DEFERRED PROOFS

A.1 Proof of Lemma C.2

Proof. Let $w_j = \mathbb{E}[p(\mathbf{x})|p(\mathbf{x}) \in I_j]$, and note that $w_j \in I_j$, a property that will be used shortly. We can write

$$\text{intCE}_B(p, J^*) = \text{width}(B) + \sum_j \Pr[p(\mathbf{x}) \in I_j]|v_j - w_j|. \tag{A.1}$$

We now bound $d(p, q_B)$ as

$$\begin{aligned}
d(p, q_B) &= \mathop{\mathbb{E}}_{\mathcal{D}^*}[|p(\mathbf{x}) - q_B(x)|] \\
&= \sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j]\, \mathbb{E}[|p(\mathbf{x}) - v_j||p(\mathbf{x}) \in I_j] \\
&\le \sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j]\, (\mathbb{E}[|p(\mathbf{x}) - w_j||p(\mathbf{x}) \in I_j] + |w_j - v_j|) \\
&\le (\sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j])\text{width}(B) + \sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j]|v_j - w_j| \\
&= \text{intCE}_B(p, J^*) \quad \text{(By equation(A.1))}
\end{aligned}$$

where the penultimate line uses the fact that conditioned on $p(\mathbf{x}) \in I_j$, $|p(\mathbf{x}) - w_j| \le \text{width}(B)$ since both values lie in the interval I_j. □

A.2 Proof of Lemma C.4

Proof. Let β be a width parameter to be chosen later. We consider the bucketing B where the first interval is $[0, b]$ for b picked randomly from the interval $[0, \beta]$. Every subsequent interval has width β (except possibly the last, which might be smaller). Denote the intervals by $I_1, \ldots, I_k$.

For the predictor q, the calibration error term for B is 0 since

$$\text{CE}_B(q) = \sum_{j\in k} |\,\mathbb{E}[\mathbf{1}(q(\mathbf{x}) \in I_j)(\mathbf{y}^* - q(\mathbf{x}))]| \le \int_{v\in[0,1]} \Pr[q(\mathbf{x}) = v]|\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x})|q(\mathbf{x}) = v]| = 0.$$

So we will try the bound the calibration term for p by comparing it to q and arguing that if they are close by, this error is small.

$$\begin{aligned}
\text{CE}_B(p, \mathcal{D}^*) &= \sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - p(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| \\
&\le \sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| + \sum_{j\in k} |\,\mathbb{E}[(q(\mathbf{x}) - p(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]|
\end{aligned} \tag{A.2}$$

We bound each of these terms separately. To bound the second term,

$$\sum_{j\in k} |\,\mathbb{E}[(q(\mathbf{x}) - p(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| = \mathbb{E}[|q(x) - p(x)|] \le \delta \tag{A.3}$$

For the first term, we have

$$\begin{aligned}\sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(p(\mathbf{x})\in I_j)]| &\le \sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(q(\mathbf{x})\in I_j)]| + \\ &\quad |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))(\mathbb{I}(p(\mathbf{x})\in I_j) - \mathbb{I}(q(\mathbf{x})\in I_j)] \\ &\le \sum_j |\mathbb{I}(p(\mathbf{x})\in I_j) - \mathbb{I}(q(\mathbf{x})\in I_j)|\end{aligned}$$

where we use $\mathrm{CE}_B(q,\mathcal{D}^*)=0$ and $|\mathbf{y}^* - q(\mathbf{x})|\le 1$. The RHS is 0 if $p(\mathbf{x})$ and $q(\mathbf{x})$ land in the same bucket, else it is 2. $p(\mathbf{x})$ and $q(\mathbf{x})$ land in different buckets if there is a bucket boundary between them, which happens with probability bounded by $|p(\mathbf{x}) - q(\mathbf{x})|/\beta$ over the random choice of b. Hence we can bound

$$\sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(p(\mathbf{x})\in I_j)]| \le \frac{2\,\mathbb{E}[|p(\mathbf{x}) - q(\mathbf{x})|]}{\beta} = \frac{2\delta}{\beta}. \tag{A.4}$$

Plugging Equations (A.3) and (A.4) back into Equation (A.2) and choosing $\beta = \sqrt{2\delta}$,

$$\begin{aligned}\mathrm{CE}_B(p,\mathcal{D}^*) &\le \delta + 2\delta/\beta. \\ \mathrm{intCE}_B(p,\mathcal{D}^*) &\le \mathrm{CE}_B(p,\mathcal{D}^*) + \mathrm{width}(B) \le \beta + \delta + 2\sqrt{\delta} \le 4\sqrt{\delta}.\end{aligned}$$

□

A.3 Proof of Lemma 4.4

Proof. The lemma can be proved by considering the level sets $\mathcal{X}_v := \{x \in \mathcal{X} : p(x) = v\}$ for $v \in [0,1]$. Within each level set, p is a constant function, and the functions $\sigma(p(x))$ formed by all choices of $\sigma : [0,1] \to A$ are all the constant functions on this level set taking value in A. Moreover, for any level set $\mathcal{X}_v$, the conditional distribution of $\mathbf{y}$ given $\mathbf{x} \in \mathcal{X}_v$ is the Bernoulli distribution with parameter $\widehat{p}(\mathbf{x})$, where $\widehat{p}(x)$ is also a constant function for $x \in \mathcal{X}_v$. Decomposing (4.4) by the level sets, the lemma follows from the definition of the best-response function $\sigma_{\mathcal{T}}$ in (4.1). □

A.4 Proof of Lemma 4.7

Proof. By Theorem 4.6,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x}),\mathbf{y}))] &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(p(\mathbf{x})) + \nabla\varphi(p(\mathbf{x}))\cdot(\mathbf{y} - p(\mathbf{x}))] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(p(\mathbf{x})) + \nabla\varphi(p(\mathbf{x}))\cdot(\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \\ &\qquad\qquad \text{(because } \mathbb{E}[\mathbf{y}|p(\mathbf{x})] = \widehat{p}(\mathbf{x})\text{)} \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x})) - \varphi(p(\mathbf{x})) - \nabla\varphi(p(\mathbf{x}))\cdot(\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))].\end{aligned}$$

This proves Equation (4.7). Similarly,

$$\mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x}),\mathbf{y}))] = \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x})\|\widehat{p}(\mathbf{x}))] = \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))].$$

Taking the difference between the two equations above, we have

$$\mathrm{CFDL}_{\mathcal{T}}(p,\mathcal{D}) = \mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x}),\mathbf{y}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x}),\mathbf{y}))] = \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))].$$

This proves Equation (4.8). □

A.5 Proof of Theorem 4.9

Proof. Let $\widehat{p}$ be the recalibration of p (Definition 6.4). By the definitions of J^* and J^p, for any $x \in \mathcal{X}$, we have

$$\mu_{J^*}(p(x)) = \widehat{p}(x), \tag{A.5}$$

$$\mu_{J^p}(p(x)) = p(x). \tag{A.6}$$

Let M denote the marginal distribution of $p(\mathbf{x})$ where $(\mathbf{x},\mathbf{y}^*) \sim \mathcal{D}$. By Lemma 4.7,

$$\begin{aligned}\mathrm{CFDL}_{\mathcal{T}}(p,\mathcal{D}) &= \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))] \\ &= \mathop{\mathbb{E}}_{\mathbf{v}\sim M}[D_{\varphi}(\mu_{J^*}(\mathbf{v})\|\mu_{J^p}(\mathbf{v}))] \\ &= D_{\varphi}(J^*\|J^p).\end{aligned}$$

□

A.6 V-shaped Divergences

In this subsection, we discuss a fundamental result about Bregman divergences (Theorem A.1) that will be used to prove Theorem 4.3.

CDL focuses on decision tasks $\mathcal{T} = (A,u)$ with $[0,1]$-bounded payoff functions $u : A \to [0,1]$. For such tasks, the corresponding convex function φ from Theorem 4.6 must have bounded subgradients:

$$\nabla\varphi(v) = u(\sigma_{\mathcal{T}}(v),1) - u(\sigma_{\mathcal{T}}(v),0) \in [-1,1] \quad \text{for every } v \in [0,1]. \tag{A.7}$$

While the convex functions φ with bounded subgradients $\nabla\varphi(v) \in [-1,1]$ form a large family, a fundamental result by [Li et al. 2022], which we include as Theorem A.1 below, shows that the divergences D_{φ} defined by this family can be captured by extremely simple functions φ that are termed *V-shaped* functions. Specifically, for each $v^* \in [0,1]$, a V-shaped function φ_{v^*} is defined as follows:

$$\varphi_{v^*}(v) = |v - v^*| \quad \text{for every } v \in [0,1].$$

The Bregman divergence $D_{\varphi_{v^*}}$ is correspondingly termed a *V-shaped* divergence, and it can be easily computed as follows: for $v_1, v_2 \in [0,1]$, we have

$$D_{\varphi_{v^*}}(v_1\|v_2) = \begin{cases} 2|v_1 - v^*| \le 2|v_1 - v_2|, & \text{if } v^* \in (v_1,v_2] \text{ or if } v^* \in (v_2,v_1]; \\ 0, & \text{otherwise.} \end{cases} \tag{A.8}$$

The following theorem gives an upper bound on the expected divergence D_{φ} for a general φ with bounded subgradient in terms of V-shaped divergences $D_{\varphi_{v^*}}$.

Theorem A.1 [Li et al. 2022]. *Let $\varphi : [0,1] \to \mathbb{R}$ be a convex function whose subgradient is bounded: $\nabla\varphi(v) \in [-1,1]$ for every $v \in [0,1]$. Then for any distribution Π of $(v_1,v_2) \in [0,1]$,*

$$\mathop{\mathbb{E}}_{(v_1,v_2)\sim\Pi} D_{\varphi}(v_1,v_2) \le \sup_{v^*\in[0,1]} \mathop{\mathbb{E}}_{(v_1,v_2)\sim\Pi} D_{\varphi_{v^*}}(v_1,v_2).$$

A.7 Relationship to ECE

We prove Theorem 4.3, which demonstrates the quadratic relationship between the CDL and the ECE. The first and last inequalities in (4.3) follow immediately from Jensen's inequality. The second inequality can be proved using the decision task $\mathcal{T}_2$ from Section 4.2. Specifically, the payoff function of $\mathcal{T}_2$ has its range bounded in $[0,1]$, so by the definition of CDL and Equation (4.6),

$$\mathrm{CDL}(p,\mathcal{D}) \geq \mathrm{CFDL}_{\mathcal{T}_2}(p,\mathcal{D}) = \mathrm{ECE}_2(p,\mathcal{D})^2.$$

Now we prove the third inequality in (4.3). By Lemma 4.7 and Theorem A.1, for any decision task $\mathcal{T}$ with $[0,1]$ bounded payoffs,

$$\begin{aligned}\mathrm{CFDL}_{\mathcal{T}}(p,\mathcal{D}) = \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))] &\leq \sup_{v^*\in[0,1]} \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi_{v^*}}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))] \\ &\leq 2\mathop{\mathbb{E}}_{\mathcal{D}}|\widehat{p}(\mathbf{x}) - p(\mathbf{x})| = \mathrm{ECE}(p,\mathcal{D}). \quad \text{(by (A.8))}\end{aligned}$$

This proves $\mathrm{CDL}(p,\mathcal{D}) \leq 2\mathrm{ECE}(p,\mathcal{D})$, as desired.

A.8 Tight examples between CDL and ECE

We prove the correctness of the two examples $(p_1,\mathcal{D}_1),(p_2,\mathcal{D}_2)$ we mentioned after Theorem 4.3 that shows the tightness of Theorem 4.3.

In the first example, we have $p_1(x) = 1/2$ and $\widehat{p}_1(x) = 1/2+\varepsilon$ for any x, so it is clear that $\mathrm{ECE}_2(p_1,\mathcal{D}_1) = \varepsilon$. To prove $\mathrm{CDL}(p_1,\mathcal{D}_1) \geq 2\varepsilon$, consider the task $\mathcal{T}_1 = (A,u)$ with two actions: $A = \{0,1\}$. The payoff function u is defined such that $u(a,y) = 1$ if $a = y$, and $u(a,y) = 0$ otherwise. The best-response function is $\sigma_{\mathcal{T}}(v) = 0$ if $v \leq 1/2$, and $\sigma_T(v) = 1$ otherwise. We have

$$\begin{aligned}\mathbb{E}[u(\sigma_{\mathcal{T}_1}(p_1(\mathbf{x})),\mathbf{y})] &= \mathbb{E}[u(0,\mathbf{y})] = \Pr[\mathbf{y}=0] = \frac{1}{2} - \varepsilon, \\ \mathbb{E}[u(\sigma_{\mathcal{T}_1}(\widehat{p}_1(\mathbf{x})),\mathbf{y})] &= \mathbb{E}[u(1,\mathbf{y})] = \Pr[\mathbf{y}=1] = \frac{1}{2} + \varepsilon.\end{aligned}$$

Taking the difference between the two expected payoffs, we get $\mathrm{CDL}(p_1,\mathcal{D}_1) \geq \mathrm{CFDL}_{\mathcal{T}_1}(p_1,\mathcal{D}_1) = 2\varepsilon$.

In the second example, we have $p_2(x) = x$ and $\widehat{p}_2(x) = x-\varepsilon$, so it is clear that $\mathrm{ECE}(p_2,\mathcal{D}_2) = \varepsilon$. Now we prove

$$\mathrm{CDL}(p_2,\mathcal{D}_2) \leq \frac{\varepsilon^2}{1-\varepsilon} = \varepsilon^2 + O(\varepsilon^3). \tag{A.9}$$

Consider any decision task $\mathcal{T} = (A,u)$ with a $[0,1]$-bounded payoff function u :

$A \to [0,1]$. By Lemma 4.7 and Theorem A.1,

$$\begin{aligned}
\mathrm{CFDL}_{\mathcal{T}}(p_2, \mathcal{D}_2) &= \mathop{\mathbb{E}}_{\mathcal{D}_2}[D_\varphi(\widehat{p}_2(\mathbf{x}) \| p_2(\mathbf{x}))] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}_2}[D_\varphi(\mathbf{x} - \varepsilon \| \mathbf{x})] \\
&\leq \sup_{v^* \in [0,1]} \mathop{\mathbb{E}}_{\mathcal{D}_2}[D_{\varphi_{v^*}}(\mathbf{x} - \varepsilon \| \mathbf{x})] \\
&= \sup_{v^* \in [0,1]} \Pr_{\mathcal{D}_2}\left[v^* - (\mathbf{x} - \varepsilon) \middle| v^* \in (\mathbf{x} - \varepsilon, \mathbf{x}]\right] \qquad \text{(by (A.8))} \\
&= \sup_{v^* \in [0,1]} \int_0^1 (v^* - (x - \varepsilon)) \mathbb{I}(v^* \in (x - \varepsilon, x]) \mathrm{d}x \\
&\leq \sup_{v^* \in [0,1]} \int_{v^*}^{v^* + \varepsilon} (v^* - (x - \varepsilon)) \mathrm{d}x \\
&= 2\varepsilon^2. \qquad \text{(A.10)}
\end{aligned}$$

Since this upper bound on the CFDL holds for any decision task $\mathcal{T}$ with a $[0,1]$-bounded payoff function, it implies (A.9), as desired.

B. THE INHERENT UNCERTAINTY IN DISTANCE TO CALIBRATION

Assume that all we know about the predictor p and distribution $\mathcal{D}^* = (\mathbf{x}, \mathbf{y}^*)$ is the distribution $J^* = (p(\mathbf{x}), \mathbf{y}^*)$. Does this specify $\mathrm{dCE}(p, D^*)$ completely? Or is there still some uncertainty on how far the closest calibrated predictor is, depending on the space $\mathcal{X}$?

We present a simple example showing that there is indeed some uncertainty. Take ϵ to be any value in $(0, 1/2)$, and let $\delta = \epsilon/(1 - 2\epsilon)$. The distribution J^* is easy to describe: $p(\mathbf{x})$ takes the values $1/2 + \delta$ and $1/2 - \delta$ each with probability $1/2$, and conditioned on each value of $p(\mathbf{x})$, $\mathbf{y}^*$ is uniformly distributed in $\{0, 1\}$.

Note that any such p is not perfectly calibrated. But it is δ far from the constant $1/2$ predictor, which is perfectly calibrated. It is easy to construct a space where this is indeed the closest calibrated predictor, so that $\mathrm{dCE}(p, \mathcal{D}^*) = \delta$.

What is perhaps less obvious is there exist spaces and predictors realizing J^* where the true distance to calibration is much smaller. We describe one such construction. Let $\mathcal{X} = \{00, 01, 10, 11\}$. Cosndier the distribution $\mathcal{D}^*$ on pairs $(\mathbf{x}, \mathbf{y}^*) \in \mathcal{X} \times \{0, 1\}$, and predictors $p_1, p_2 : \mathcal{X} \to [0, 1]$ given below:

x	$\Pr_{\mathcal{D}^*}[\mathbf{x} = x]$	$\mathbb{E}_{\mathcal{D}^*}[\mathbf{y}^* \mid \mathbf{x} = x]$	$p_1(x)$	$p_2(x)$
00	$\frac{1}{2} - \epsilon$	$\frac{1}{2} - \delta$	$\frac{1}{2} - \delta$	$\frac{1}{2} - \delta$
01	ϵ	1	$\frac{1}{2} - \delta$	$\frac{1}{2}$
10	ϵ	0	$\frac{1}{2} + \delta$	$\frac{1}{2}$
11	$\frac{1}{2} - \epsilon$	$\frac{1}{2} + \delta$	$\frac{1}{2} + \delta$	$\frac{1}{2} + \delta$

The predictor p_1 is not perfectly calibrated, indeed we have chosen δ such that the joint distribution of $(p_1(\mathbf{x}), \mathbf{y}^*)$ is exactly J^*: conditioned on either prediction value in $\{1/2 \pm \delta\}$, the bit $\mathbf{y}^*$ is uniformly random. In contrast, the predictor p_2 is easily seen to be calibrated.

Observe that p_1 and p_2 agree on 00 and 11. They disagree by δ on 01 and 10, which each have ϵ probability under $\mathcal{D}^*$, so $d(p_1, p_2) = 2\epsilon\delta = \Theta(\epsilon^2)$. This establishes the difficulty of pinning down the true distance to calibration within a quadratic factor.

C. RELATING UPPER AND LOWER DISTANCES TO CALIBRATION

In this section, we prove Theorem 6.9 showing that the upper and lower distance to calibration can be at most quadratically far apart. This shows that the simple example in Appendix B is nearly tight. We follow the proof strategy of [Blasiok et al. 2023a] using the notion of *interval calibration error*.

C.1 Interval Calibration Error

Definition C.1 *Interval Calibration Error [Blasiok et al. 2023a].* A interval partition B is a partition of the interval $[0, 1]$ into disjoint intervals $I_1, \ldots, I_k$. We let the width of the partition width(B) be the length of longest interval. Given a predictor p, we define its calibration error and interval calibration error for B respectively as

$$\mathrm{CE}_B(p, \mathcal{D}^*) = \sum_{j \in [k]} |\, \mathbb{E}[(\mathbf{y}^* - p(\mathbf{x}))\mathbf{1}(p(\mathbf{x}) \in I_j)]|$$
$$\mathrm{intCE}_B(p, \mathcal{D}^*) = \mathrm{CE}_B(p, J^*) + \mathrm{width}(B).$$

The *interval calibration error* minimizes over all interval partitions B:

$$\mathrm{intCE}(p, \mathcal{D}^*) = \min_B \mathrm{intCE}_B(p, \mathcal{D}^*).$$

The definition of intCE_B involves two terms that represent a tradeoff: the calibration error term, and the width term that penalizes partitions which use large width intervals. Intuitively, as the intervals grow larger it is easier to reduce calibration error, since we are allowed to cancel out the point-wise errors $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] - p(\mathbf{x})$ over larger intervals; but the width penalty also grows larger. At one extreme, we can think of the width 0 case as corresponding to the ECE. At the other extreme, by taking the single interval $[0, 1]$, we pay $\mathbb{E}[\mathbf{y}^* - p(\mathbf{x})]$ which is 0 if the expectations of $\mathbf{y}^*$ and $p(\mathbf{x})$ are equal; a very weak calibration guarantee. But now the width penalty is 1.

Formal justification for the definition comes from the following observation. The canonical predictor q_B for an interval partition B and a distribution $\mathcal{D}^*$ is the predictor where for all $x \in I_j$, the q_B predicts $v_j = \mathbb{E}[y^*|p(\mathbf{x}) \in I_j|$. It is easy to see that q_B is perfectly calibrated for $\mathcal{D}^*$.

Lemma C.2. *The canonical predictor q_B for $B, \mathcal{D}^*$ satisfies $d(p, q_B) \leq \mathrm{intCE}_B(p, \mathcal{D}^*)$.*

This leads to the following upper bound:

Theorem C.3. *[Blasiok et al. 2023a] We have $\overline{\mathrm{dCE}}(p, \mathcal{D}^*) \leq \mathrm{intCE}(p, \mathcal{D}^*)$.*

To prove Theorem C.3 we observe that the canonical predictor q_B can be viewed as a post-processing of the predictor p, since we can write $q_B(x) = \kappa(p(x))$ where $\kappa(t) = v_j$ for $t \in I_j$. Thus by Lemma C.2,

$$\overline{\mathrm{dCE}}(p, \mathcal{D}^*) \leq d(p, q_B) \leq \mathrm{intCE}_B(p, \mathcal{D}^*).$$

Minimizing over all B completes the proof.

The reader might wonder, why define yet another calibration measure? The answer is two-fold:

—Interval calibration error gives a simple yet powerful upper bound on the upper distance to calibration. In the next subsection, this allows us to relate the upper and lower distance to calibration, showing that they are never more than quadratically far apart. This is formally proved in Theorem 6.9, showing the gap example in Corollary 6.8 is the worst possible (up to constants).

—It presents a rigorous alternative to heuristic measures like bucketed ECE: regularize the calibration error by adding the max bucket width. This allows for meaningful comparison of calibration scores obtained using different number or other choice of buckets, rather than leaving the number of buckets as a hyperparameter.

C.2 Proof of Theorem 6.9

Let us pick $\mathcal{X}, \mathcal{D}^*, p$ to be the space, distribution and predictor respectively that achieve the lower distance to calibration for J^*. So there exists a perfectly calibrated predictor $q : \mathcal{X} \to [0,1]$ such that $d(p,q) = \underline{\mathrm{dCE}}(p, \mathcal{D}^*) = \delta$. We wish to infer the existence of a bucketing B so that $\mathrm{intCE}_B(p, \mathcal{D}^*)$ is small. By Theorem C.3, this will imply that the upper distance is bounded. Corollary 6.8 tells us that we cannot hope for an upper bound better than $\sqrt{\delta}/2$. It turns out that this is not far from the best possible (see Section A.2 for proof):

LEMMA C.4. *There exists a bucketing B such that* $\mathrm{intCE}_B(p, \mathcal{D}^*) \le 4\sqrt{\delta}$.

Combining this lemma with Corollary 6.8, we have completed the proof:

$$\overline{\mathrm{dCE}}(p, \mathcal{D}^*) \le \mathrm{intCE}(p, \mathcal{D}^*) \le \mathrm{intCE}_B(p, \mathcal{D}^*) \le 4\sqrt{\delta} = 4\sqrt{\underline{\mathrm{dCE}}(p, \mathcal{D}^*)}.$$

Algorithmic Delegated Choice: An Annotated Reading List

MOHAMMAD T. HAJIAGHAYI
University of Maryland
and
SUHO SHIN
University of Maryland

The problem of delegated choice has been of long interest in economics and recently on computer science. We overview a list of papers on delegated choice problem, from classic works to recent papers with algorithmic perspectives.

Categories and Subject Descriptors: B.6.3 [**Theory of computing**]: Algorithmic mechanism design

General Terms: Mechanism Design,

Additional Key Words and Phrases: Delegation

1. INTRODUCTION

The *delegated choice* problem is a fundamental model of principal-agent interaction with numerous real-world applications, capturing the tension when a decision maker (principal) delegates the role of decision making to an informed but self-interested agent. The model has its roots in classic economic theory introduced by [Holmström 1978], and has since evolved into a rich interdisciplinary area spanning economics, computer science, and operations research. It considers a scenario where the principal *cannot* commit to a contingent monetary transfer, and thus the principal needs to commit to a mechanism that specifies the characteristics of the agent's proposals that she is willing to accept. Such a model is particularly motivated by practical scenarios such as public regulators who can only accept or reject proposals from private sector, or investors relying on recommendations from financial advisors who are not contractually incentivized by the investor's returns.

In its canonical form, the principal must select an action from a discrete (or often continuous) set Ω, where each action $i \in \Omega$ has a pair of random utility values (X_i, Y_i): one for the principal and one for the agent. Only the agent observes these values and proposes an action for selection, while the principal only knows the distribution from which these random values are drawn, introducing an information asymmetry. Once the agent observes the realizations, he sends a signal (*e.g.,* proposing an action) to the principal, who then makes the final decision. The agent seeks to maximize his own utility rather than the principal's, resulting in moral hazard. To mitigate this, the principal commits to a *screening mechanism* that selectively accepts the proposed action based on its values. For instance, an

Authors' addresses: hajiagha@umd.edu, suhoshin@umd.edu

eligible set $E_i \subseteq \mathbb{R}^2$ can be announced to the agent so that the principal will only accept the proposed action i if $(X_i, Y_i) \in E_i$.

Recent work has shifted from classical results on existence and structure of optimal mechanisms to more algorithmic and computational perspectives. This reading list collects key papers in this growing literature, highlighting recent developments in various aspects of the delegated choice problem, including its connections to the prophet inequality [Krengel and Sucheston 1977], Pandora's box [Weitzman 1978], and broader stochastic optimization problems.

Our goal is to introduce the core problem setup, present both classic and recent contributions, and illustrate how this line of work connects to various research areas in the EC community. As such, this article is neither exhaustive nor comprehensive; we hope it will serve as a useful starting point for readers to grasp the central ideas and emerging directions in delegated choice problem.

(1) Bengt Rober Holmström. On incentives and control in organizations, *Ph.D. dissertation thesis*, 1978 & Bengt Rober Holmström, On the theory of delegation, *Bayesian Models in Economic Theory*, 1984.

Seminal work by [Holmström 1978; 1984] provides the *foundational framework* for delegation as an optimal screening mechanism. Instead of the discrete choice model described above, the principal delegates an optimization problem by choosing a screening set $A' \subseteq A$ from which the agent selects an action $a \in A'$ that yields payoffs $v(a, \theta)$ and $u(a, \theta)$ to the principal and agent, respectively, where θ is a realized state observed only by the agent. Holmstrom characterizes conditions under which a single interval is optimal, and shows that the discretion given to the agent increases as his preferences align more closely with the principal's.

(2) Mark Armstrong, John Vickers. A model of delegated project choice, *Econometrica*, 2010.

[Armstrong and Vickers 2010] provides the first *discrete choice* model, which serves as a foundation for subsequent works. There are n available actions,[1] each with a type (u, v) drawn i.i.d. from a distribution, where the agent receives payoff v and the principal receives $v + \alpha u$ for some $\alpha \geq 0$. Only the proposed action's type is verifiable, so the principal's goal is to design a screening mechanism over admissible types (u, v), and the authors identify optimal mechanisms under specific payoff assumptions.

(3) Jon Kleinberg, Robert Kleinberg. Delegated search approximates efficient search, *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, 2018.

[Kleinberg and Kleinberg 2018] studies the approximate efficiency of the optimal delegation mechanism with respect to the *first-best* benchmark, where the principal observes the utilities of all actions in hindsight and can choose what

[1]In fact, they consider a setting where n is also drawn from a known distribution.

she wants, building on the model of [Armstrong and Vickers 2010].[2] It considers precisely the problem setup described in the introduction, and uncovers a surprising connection to the classical *prophet inequality* [Krengel and Sucheston 1977]. Specifically, they prove an *equivalence* between the delegated choice problem and a version of prophet inequality with oblivious stopping rules. This leads to several delegation gap bounds, including a 1/2-approximation with a threshold mechanism in the general case and a $(1 - 1/e)$-approximation in the i.i.d. case.

(4) Kiarash Banihashem, Mohammad T. Hajiaghayi, Piotr Krysta, Suho Shin. Delegated choice with combinatorial constraints, *Proceedings of the 2025 ACM Conference on Economics and Computation (EC)*, 2025.

[Banihashem et al. 2025] considers a natural follow-up to [Kleinberg and Kleinberg 2018], asking to what extent the connection between delegated choice and prophet inequalities carries over. They study a multi-choice setting with a universe U of n actions and a family of feasible sets $\mathcal{I} \subset 2^U$, where the principal aims to select a set $I \in \mathcal{I}$ to maximize $X_I = \sum_{i \in I} X_i$. Notably, they provide the first *provable separation* between the two problems by showing that delegated choice under downward-closed constraints allows constant-factor approximation, whereas prophet inequalities does not [Rubinstein 2016]; they further show that the correspondence between the two problems holds *if and only if* the constraint is a matroid.

(5) Curtis Bechtel, Shaddin Dughmi. Delegated stochastic probing, *12th Innovations in Theoretical Computer Science (ITCS)*, 2021.

[Bechtel and Dughmi 2021] proposes a delegated stochastic probing problem where the agent probes actions under an outer constraint $\mathcal{I}_{out}$ and proposes a feasible set under an inner constraint $\mathcal{I}_{in}$. Without an outer constraint or probing cost, their model coincides with [Banihashem et al. 2025], and they show that a greedy prophet inequality strategy against an almighty adversary, who observes every random bit of the algorithms and environments and decides a worst-case instance, can be implemented in this setting. This yields several immediate corollaries on matroid, matching, and knapsack constraints via greedy online contention resolution schemes [Feldman et al. 2016].[3]

(6) Ali Kohdabakhsh, Emmanouil Pountourakis, Samuel Taggart. Simple delegated choice, *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2024.

The assumption that the proposed action's utilities are easily verifiable is admissible as misreporting could be verifiable by implementing the action or it

[2]They also consider a search problem where the agent undergoes a costly search process to seek solutions, with connection to Pandora's box problem [Weitzman 1978].

[3]We note here that this does not contradict the necessity of matroid results shown in [Banihashem et al. 2025], since [Bechtel and Dughmi 2021] guarantees only that the principal's utility exceeds the prophet's, not an exact equivalence between the agent's choice and the prophet inequality algorithm's decision as shown by [Banihashem et al. 2025].

may incur a reputation effect. On the other hand, this could often be impractical when the delegation happens as a one-off interaction or if it entails an expensive cost of verification such as the delegated decision of a governmental policy. [Khodabakhsh et al. 2024] considers a mechanism that *does not depend* on the utilities of the proposed values, but the principal completely rules actions in or out based on the distributional knowledge. They show that competing with the first-best benchmark is hopeless, and that the problem of computing the optimal mechanism is NP-hard, which is complemented by their 1/3 approximate deterministic mechanism.

(7) Mohammad T. Hajiaghayi, Piotr Krysta, Mohammad Mahdavi, Suho Shin. Delegation with costly inspection, *Proceedings of the 2025 ACM Conference on Economics and Computation (EC)*, 2025.

[Hajiaghayi et al. 2025] directly addresses the verifiability assumption by allowing the principal to *inspect* the proposed action, and possibly others, at deterministic costs c_i, to verify utilities. In their extension of [Kleinberg and Kleinberg 2018], the agent may misreport if inspection is unlikely, and delegation itself incurs a fixed cost c_{del}. This model generalizes the Pandora's box problem with nonobligatory inspection [Doval 2018], inheriting its NP-hardness [Fu et al. 2023; Beyhaghi and Cai 2023], and they show that while the first-best benchmark cannot be approximated, constant-factor approximate mechanisms exist in both costless and costly delegation settings when the cost of delegation is high or low.

(8) Suho Shin, Keivan Rezaei, Mohammad T. Hajiaghayi. Delegating to multiple agents, *Proceedings of the 2023 ACM Conference on Economics and Computation (EC)*, 2023.

While the preceding works focus on Bayesian mechanisms, a few have explored prior-independent mechanisms in relaxed settings.[4] [Shin et al. 2023] studies a multi-agent delegated choice problem where each agent proposes an action, but only the selected agent receives nonzero utility, introducing competition that benefits the principal. They consider both Bayesian and prior-independent mechanisms and show that a constant-factor prior-independent mechanism exists in the complete information setting with symmetric agents, whereas the benefit of having multiple agents depends heavily on the agents' information and symmetry.

(9) Curtis Bechtel, Shaddin Dughmi. Efficient multi-agent delegated search, *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2025.

[Bechtel and Dughmi 2025] improves the results for Bayesian mechanisms under the incomplete information setting introduced by [Shin et al. 2023]. They achieve an approximation factor tending to 1 as the number of agents increases,

[4]In the standard setup with a single agent, one can easily see that no prior-independent mechanism can approximate the first-best benchmark.

when the agents have symmetric sets of actions that are not necessarily i.i.d. This strengthens the approximation factor and relaxes the constraints introduced by [Shin et al. 2023].

(10) Mohammad T. Hajiaghayi, Mohammad Mahdavi, Keivan Rezaei, Suho Shin. Regret analysis of repeated delegated choice, *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Distributional knowledge, in practice, is usually constructed from historical data. [Hajiaghayi et al. 2024] considers a variant with prior-independent mechanisms where the principal can repeatedly interact with the agent to construct estimates of the distributions. They frame their setup as a stochastic multi-armed bandit problem, and propose no-regret learning algorithms for myopic and farsighted agents under the Lipschitzness assumption of the utilities, using tools from bandits with perturbed outputs.

REFERENCES

Armstrong, M. and Vickers, J. 2010. A model of delegated project choice. *Econometrica 78,* 1, 213–244.

Banihashem, K., Hajiaghayi, M. T., Krysta, P., and Shin, S. 2025. Delegation with costly inspection. *arXiv preprint arXiv:2506.07162*.

Bechtel, C. and Dughmi, S. 2021. Delegated stochastic probing. *Innovations in Theoretical Computer Science (ITCS)*.

Bechtel, C. and Dughmi, S. 2025. Efficient multi-agent delegated search. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2431–2433.

Beyhaghi, H. and Cai, L. 2023. Pandora's problem with nonobligatory inspection: Optimal structure and a ptas. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 803–816.

Doval, L. 2018. Whether or not to open pandora's box. *Journal of Economic Theory 175*, 127–158.

Feldman, M., Svensson, O., and Zenklusen, R. 2016. Online contention resolution schemes. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 1014–1033.

Fu, H., Li, J., and Liu, D. 2023. Pandora box problem with nonobligatory inspection: Hardness and approximation scheme. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 789–802.

Hajiaghayi, M., Mahdavi, M., Rezaei, K., and Shin, S. 2024. Regret analysis of repeated delegated choice. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 9757–9764.

Hajiaghayi, M. T., Krysta, P., Mahdavi, M., and Shin, S. 2025. Delegation with costly inspection. *Available at SSRN 5284400*.

Holmström, B. R. 1978. *On Incentives and Control in Organizations*. Stanford University.

Holmström, B. R. 1984. On the theory of delegation. *Bayesian Models in Economic Theory*, 115–141.

Khodabakhsh, A., Pountourakis, E., and Taggart, S. 2024. Simple delegated choice. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 569–590.

Kleinberg, J. and Kleinberg, R. 2018. Delegated search approximates efficient search. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 287–302.

Krengel, U. and Sucheston, L. 1977. Semiamarts and finite values.

Rubinstein, A. 2016. Beyond matroids: Secretary problem and prophet inequality with general constraints. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 324–332.

Shin, S., Rezaei, K., and Hajiaghayi, M. 2023. Delegating to multiple agents. In *Proceedings of the 24th ACM Conference on Economics and Computation*. 1081–1126.

Weitzman, M. 1978. *Optimal search for the best alternative*. Vol. 78. Department of Energy.