UNIVERSITY OF CALIFORNIA

Los Angeles

On the Societal Impact of Human-Technology Interaction

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Management

by

Jian Gao

2024

ABSTRACT OF THE DISSERTATION

On the Societal Impact of
Human-Technology Interaction

by

Jian Gao
Doctor of Philosophy in Management
University of California, Los Angeles, 2024
Professor Francisco Castro Altamirano, Chair

As technology continues to advance rapidly, understanding its broader societal implications becomes increasingly important. This thesis explores the intersection between human and emerging technologies—specifically generative artificial intelligence (AI), autonomous vehicles (AVs), and hybrid marketplaces—and their potential impacts on society. Through three interconnected studies, we investigate how these technologies influence user behavior, market dynamics, and service quality, providing valuable insights for policymakers.

In Chapter 2, we delve into the interaction between humans and generative AI. While generative AI can boost productivity, the content it produces may not always align with user preferences. To study this effect, we introduce a Bayesian framework where heterogeneous users decide how much information to share with the AI, balancing a trade-off between output fidelity and communication cost. We reveal that these interactions can lead to societal challenges such as homogenization and bias. Our findings highlight the risk of reduced diversity in outputs, especially when AI-generated content is used to train the next generation of AI systems, potentially resulting in a "homogenization death spiral." We also assess the

impact of AI bias, demonstrating how AI biases can lead to societal bias. Importantly, we suggest that facilitating human-AI interactions can mitigate these risks.

Chapter 3 investigates the societal effects of introducing AVs into a ride-hailing market which is currently served by human drivers (HVs). We develop a game-theoretical queueing model in which a platform aims to maximize its profit while HVs make strategic joining decisions. Our analysis indicates that incorporating AVs may degrade service levels, as the platform may prioritize AVs, negatively impacting HVs' earnings and driving them out of the market. We then reveal that this reduction in service level is not uniform in a city: high-demand areas, such as downtown areas, may maintain reasonable service levels, while remote areas may experience a large decline in service level. Then, using New York City data, we build a highly detailed simulation of the operations of a ride-hailing platform to further validate our theoretical model in a more realistic setting and demonstrate the additional effects on service levels. This study underscores the importance of balancing profitability with service quality when introducing AVs in the transportation sector.

In Chapter 4, we extend the analysis of Chapter 3 and focus on the strategic decisions of profit-maximizing firms operating in "hybrid marketplaces" consisting of both private and flexible supply agents. The firm can decide the number of private agents to employ, paying them regardless of their work, while flexible agents make their own revenue and pay a commission to the firm. We develop a general framework for supply prioritization, applicable to any firm using a mix of employees (private agents) and contractors (flexible agents), and capable of handling complex supply management policies. Our findings show that without prioritization, using hybrid supply is not optimal. However, effective prioritization strategies can enhance profitability by increasing the productivity of private agents, albeit at the cost of reducing flexible supply participation. Therefore, the firm tends to prioritize private supply in "over-supplied" markets, but may prioritize flexible supply in "under-supplied" markets, where a slight increase of supply increases can significantly impact outcomes. These insights highlight the critical role of prioritization in managing hybrid supply in markets.

The dissertation of Jian Gao is approved.

Charles J. Corbett

Christopher Siu Tang

Sébastien Martin

Francisco Castro Altamirano, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisors, Professors Francisco Castro and Sébastien Martin, for their invaluable guidance, support, and encouragement throughout my doctoral studies. Francisco has been my guiding light throughout this journey. We spent countless hours discussing models, solving proofs, and refining stories. His rigor and high standards for research continuously encouraged me to become a better researcher. He always made time to help me whenever I needed it. His patience and energy have been my anchor from the beginning of my studies to the challenging job search period. The time spent in his office, working together on the whiteboard, will forever be cherished memories. Sébastien's boundless energy and passion have consistently inspired me. His insightful suggestions have profoundly shaped my research. Whenever progress differed from what we expected, Sébastien was always able to find the silver lining and redirect our efforts toward success. His continuous stream of innovative ideas and his talent for identifying important problems, crafting compelling stories, and pushing research forward have taught me immensely. The only regret I have is the physical distance and limited in-person meetings over the past years. However, these never hindered our numerous meetings, conversations, and discussions. Francisco and Sébastien have taught me how to identify research questions I am passionate about, delve deeper into research problems, and become a good presenter. Our diverse styles, traits, and specializations made us, in my view, the best team in the world. I am also deeply grateful for their understanding and sincere advice, not only in research but also in career and life decisions. I cannot express enough gratitude to them. Without their mentorship, I would not have become who I am today. I feel really fortunate to have been their student and to have learned so much from them over the years.

I am also profoundly grateful to the esteemed members of my committee, Professors Christopher Tang and Charles Corbett, for their time, expertise, and insightful suggestions. Their diverse perspectives and rigorous standards have significantly enriched this work. I

am really thankful to Chris for his guidance in both career and life decisions. I always try to follow in his steps and hope to become someone like him one day: knowledgeable, intelligent, energetic, and meticulous. His passion has motivated me to pursue higher goals. I have been honored to be his teaching assistant in his classes, where I learned a lot about how to be a great thinker, presenter, and teacher. I would also like to extend my heartfelt thanks to Charles, for his critical feedback on my research and advice on my career. His encouragement has been invaluable, inspiring me to stay optimistic and keep moving forward. Additionally, I am thankful to other faculty members and staff at DOTM, especially Professors Reza Ahmadi, Fernanda Bravo, Felipe Caro, Velibor Mišić, Kumar Rajaram, and Scott Rodilitz for their help and suggestions throughout my studies. I would also like to thank Craig Jessen and Karina Morales for their support in managing the logistics over the past years.

I am fortunate to have had so many great people share my journey during my time at Anderson. I am deeply thankful to my peers at DOTM: İrem Akchen, Yi-Chun Akchen, Yizhuo Dong, Saeed Ghodsi, Martin Gonzalez Cabello, Xinyi Guan, Jingyuan Hu, Muzhi Ma, Nareen Molugu, Zach Siegel, Zhiyuan Sun, Abolfazl Taghavi, Mirel Yavuz, and Jingwei Zhang. Thank you all for your tremendous help and support over the years. It was wonderful to have you by my side, sharing experiences and feelings throughout this journey. While it was unfortunate that we could not spend more time together in person due to the COVID-19 pandemic, it only made the time we did share even more precious. I would not be where I am today without each of you.

Outside of Anderson, I also want to express my deep appreciation to my manager, Garrett van Ryzin, my mentor, Lee Dicker, and my colleagues, Sami Serkan Özarık and Yash Kanoria, during my internship at Amazon. Their mentorship and support provided me with invaluable professional experience and personal growth. I was fortunate to work with them on amazing projects, and I thoroughly enjoyed seeing ideas evolve from theories to practical applications. Garrett, thank you for your advice on projects and career guidance. From you, I learned how to translate theoretical knowledge into impactful and practical outcomes. Sami, thank

you for working with me, patiently explaining everything when I was new, and sharing your PhD experiences. Your patience and passion strongly motivated me to continue my journey.

I am eternally grateful to my parents for their unconditional love, support, and sacrifices. I am thankful to them for giving me life and growing me up so that I can see this colorful world and have such a wonderful life. Their belief in me has been the foundation of all my achievements. While I was young and stubborn sometimes, they were always there whenever I needed them. I am especially appreciative of my mom, Deling, who dedicated so much energy and many years to educating and caring for me. She is the strongest and most beautiful mother in the world. I hope my parents can be healthy and happy forever.

Finally, I would like to thank my fiancée, Feiran, for her unwavering love, patience, and understanding. She has been my greatest source of strength. Her support has been my backing through the highs and lows of this journey. Because of her, I always feel confident and fearless, no matter what happens. It was challenging to maintain a long-distance relationship for years, but it will soon come to an end. I eagerly look forward to starting a new chapter of our life together.

VITA

EDUCATION

2017 – 2019 **University of Toronto**
Master of Applied Science (Operations Research)

2012 – 2017 **University of Toronto**
Honors Bachelor of Science (Economics and Mathematics)

PROFESSIONAL EXPERIENCE

2022 Research Scientist Intern, Amazon

HONORS AND AWARDS

2019 – 2024 Anderson Fellowship, UCLA Anderson School of Management

2021 Easton Technology Management Center Grant, UCLA

PUBLICATIONS

PUBLISHED & UNDER REVIEW PAPERS

Francisco Castro, **Jian Gao**, Sébastien Martin, "Human-AI Interactions and Societal Pitfalls," *Proceedings of the 25th ACM Conference on Economics and Computation*, 2024.

Francisco Castro, **Jian Gao**, Sébastien Martin, "Autonomous Vehicles in Ride-Hailing and the Threat of Spatial Inequalities," under review at *Manufacturing & Service Operations Management*, 2024.

Francisco Castro, **Jian Gao**, Sébastien Martin, "Supply Prioritization in Hybrid Marketplaces," working paper, 2024.

CHAPTER 1

Introduction

In recent years, machine learning and *artificial intelligence* (AI) have propelled advancements in autonomous technologies that present new opportunities that enhance productivity by assisting or even replacing human labor. However, they also present a lot of challenges for society. For example, *generative AIs* can aid in tasks like writing, coding, and designing; and *autonomous vehicles* (AV) are expected to improve the efficiency of our transportation system with reduced costs. However, the individual and societal implications of these technologies may be more intricate, encompassing potential risks and raising critical questions that involve social welfare and fairness. In this context, this thesis aims to comprehend the multifaceted impacts that emerge from the interaction between humans and these new technologies.

Specifically, our goal is to provide valuable insights into autonomous technologies for policymakers and other stakeholders by integrating principles from operations research, economics, statistics, information theory, and computer science. Thus far, this thesis has focused on (1) modeling complex real-world systems and the strategic behavior of relevant stakeholders in the context of employing autonomous technologies; (2) and better understanding how strategic decisions impact all stakeholders and society at large. We also constructed simulations that more closely mirror the realistic system to support and extend the theoretical results from the model.

This thesis is organized as follows. Chapter 2 discusses a timely topic regarding generative AIs: the potential societal consequences when users interact with AI systems. To this end, we develop a parsimonious Bayesian model in which users with heterogeneous preference

rationally decide how much information to share with the AI, balancing output fidelity and communication costs. In Chapter 3, we present a game-theoretical queueing model and a numerical simulation to understand the optimal strategy for a ride-hailing platform to manage a hybrid fleet composed of human-driven vehicles (HV) and AVs. This chapter aims to assess the aggregate-level consequences of introducing AVs in a city in terms of service levels. Chapter 4 generalizes the previous analysis by examining how a profit-maximizing firm prioritizes its supply in a "hybrid marketplace" comprising both private and flexible supply agents. We conclude this thesis in Chapter 5. Detailed proofs and additional numerical results are provided in the appendices for further reference.

In the following sections, we will outline the main ideas explored in each chapter.

1.1 Human-AI Interactions and Societal Pitfalls

As people start using generative AI to become more productive, this new work paradigm may also lead to undesirable side effects. In particular, the boost in productivity may come at the expense of users' idiosyncrasies, such as personal style, tastes, and preferences that we would naturally express without AI. To let users express their preferences, many AI systems (e.g., ChatGPT) allow users to interact with AIs, and users can review and edit the AI-generated output themselves. However, aligning a user's intentions with an AI's output can take time. For example, when generating a picture with an AI, it could be difficult to describe everything about colors, shapes, and styles by articulating more detailed prompts to better align the AI output with our preferences. Thus, there exists a trade-off between AI output fidelity and communication cost.

To study the societal effect of this interaction, we introduce a Bayesian framework in which each rational user can exchange information with the AI to align its output with their heterogeneous preferences. The AI has a knowledge of the distribution of preferences in the population and uses a Bayesian update to create an output with maximal expected fidelity

given the information shared by the user. Users choose the amount of information they share to maximize their utility, balancing the communication cost with the output fidelity.

We show that the interplay between these individual-level decisions and AI training may lead to societal challenges: homogenization and bias. First, we prove that the AI-generated output distribution has a lower variance than the users' preference distribution. This phenomenon is exacerbated when AI-generated content is used to train the next generation of AI: we show numerically that the users' rational decisions and the AI's training process can mutually reinforce each other, leading to a homogenization "death spiral". Additionally, we also study the effects of AI bias, identifying who benefits or loses when using an AI model that does not accurately reflect the population preference distribution. We show that the censoring type of bias (e.g., biasing against the more unique preferences) only slightly improves the utility of common-preference users but significantly reduces the utility of unique users. Directional biases (e.g., a slightly left-leaning AI) are detrimental to users whose preferences are opposite to the AI bias, leading to a societal bias.

Nonetheless, our research also demonstrates that creating models that improve human-AI interactions can significantly limit these risks and preserve population diversity. For example, a clearer interface and guidance about how to articulate a good prompt can incentivize users to provide more information, thereby better matching their original preferences.

1.2 Autonomous Vehicles in Ride-Hailing and the Threat of Spatial Inequalities

When ride-hailing platforms start to adopt self-driving cars, they will likely have to manage a mixed fleet of HVs and AVs. The purpose of this work is to describe some potential challenges that could arise for a ride-hailing platform operating a mixed fleet of HVs and AVs. Because the two types of vehicles have very different costs for the platform, the platform will need to design cost-effective dispatch strategies to maximize its profit. We want to understand

these strategies and their impact on key performance metrics such as the service level and the spatial equality of access to transportation.

To this end, we develop a game-theoretical queueing model in which a revenue maximizer platform and vehicles interact in two stages. In the first stage, given a certain number of AVs, humans drivers decide to join the platform by gauging their earning rates against an outside option. The number of HVs is therefore decided by a wage equilibrium. In the second stage, the platform organizes two queues—one for each vehicle type—and decides how to distribute new requests between the vehicles in the queues in order to maximize its profit rate. This two-stage setting captures that in practice, drivers make joining decisions on a longer time scale than platform's matching decisions, which are made much more frequently.

Our results demonstrate that the introduction of AVs may deteriorate the service level. As the platform incorporates AVs, it may prioritize them to maximize profit, which affects the earnings of HVs and drives them out of the market. More precisely, we find that when the platform uses the optimal allocation policy, the presence of AVs decreases the earning of HVs and therefore expels them from the market. This substitution is disproportionate in that one additional unit of AVs leads to more than one unit of HVs leaving the platform, which, in turn, causes an overall service level decline.

We then reveal that the reduction of service level is not homogeneous across areas in a city: while the more profitable high-demand areas, such as downtown areas, may see a high concentration of vehicles and reasonable service levels, remote locations may suffer from a drop in service level. The reason behind it is that a revenue-maximizing platform will try to distribute AVs in the areas with higher demand, which will in turn push the HVs to lower-demand areas. As a result, the earning rate of HVs is reduced not only due to the additional number of AVs in the market, but also because they end up in less profitable areas.

Using New York City data, we build a simulation that more closely resembles the operations of a ride-hailing platform. For instance, each vehicle is simulated individually within the city network. The simulation also assumes that customers have a utility model to choose

between requesting an AV or an HV, while the platform can influence their choices through a differentiated pricing strategy. We confirm that our theoretical results from the queueing model still hold. Urban areas and airports reap most of the benefits of the introduction of AVs, while suburban areas experience low service levels. We also demonstrate additional effects on service levels that also lead to spatial inequality, such as distance between areas and demand imbalance.

1.3 Supply Prioritization in Hybrid Marketplaces

In this paper, we develop a general framework to study how a profit-maximizing firm prioritizes its supply in a "hybrid marketplace" composed of both private and flexible supply agents. The firm can choose how many private agents to operate and pay them regardless of their work. Flexible agents, on the other hand, make their own revenue, pay a commission rate to the firm, and enter or exit the marketplace in equilibrium. For example, private agents can be AVs or employees, and flexible agents can be HVs or contractors. Instead of focusing on any complex pricing/matching/supply management policies, we try to discern general principles and effects through the revenue change that supply prioritization implies for all stakeholders. For instance, prioritizing private agents would yield more revenue from private agents than when agents are treated equally.

Our main results show that prioritization strategies can make it optimal for the firm to operate a hybrid marketplace, even if the private supply is costly. Prioritizing private supply makes private agents particularly productive—justifying the firm's investment—but comes at the expense of flexible supply that may flee the market. Hence, the firm prioritizes private supply when the market is "over-supplied," but it may also prioritize flexible supply in "under-supplied" markets in which a slight increase in supply can have a significant impact. Our results provide a general understanding of the advantages of hybrid marketplaces and the key role of supply prioritization.

CHAPTER 2

Human-AI Interactions and Societal Pitfalls

2.1 Introduction

Generative artificial intelligence (AI) systems, particularly large language models (LLMs), have improved at a rapid pace. For example, ChatGPT recently showcased its advanced capacity to perform complex tasks and human-like behaviors (OpenAI, 2023b), reaching 100 million users within two months of its 2022 launch (Hu, 2023). This progress is not limited to text generation, as demonstrated by other recent generative AI systems such as Midjourney (Midjourney, 2023) (a text-to-image generative AI) and GitHub Copilot (Github, 2023) (an AI pair programmer that can autocomplete code). Eloundou et al. (2023) estimated that about 80% of the U.S. workforce could be affected by the introduction of LLMs, and 19% of the workers may have at least 50% of their tasks impacted. In particular, AI can make users more productive by generating complex content in seconds, while users can simply communicate their preferences. For example, Noy and Zhang (2023) highlighted that ChatGPT can substantially improve productivity in writing tasks, and GitHub claims that Copilot increases developer productivity by up to 55% (Kalliamvakou, 2023).

However, content generated with the help of AI is not exactly the same as content generated without AI. The boost in productivity may come at the expense of users' idiosyncrasies, such as personal style and tastes, preferences we would naturally express without AI. To let users express their preferences, many AI systems let users edit their prompt (e.g., Midjourney) or allow more natural interactions (e.g., ChatGPT), and users can always review and

edit the AI-generated output themselves (Vaithilingam et al., 2022). However, aligning a user's intentions with an AI's output can take time and may not always be worth it if the AI's first or default output "does the job." Consider a simple example where we use Copilot to code a Python function that calculates the sum of numbers in a nested list. Figure 2.1 shows that Copilot's default output (the first to the left) was correct and functional. However, it did not correspond to our own way of writing the same function given enough time (at the bottom of the figure). To push Copilot to better match our style, we could provide more information by articulating a more detailed prompt. However, the figure shows this may require many steps, which goes against the goal of being more productive. Similarly, Lingard (2023) described guiding ChatGPT through incremental prompting. In essence, users' time and effort to convey information about their desired outcome to an AI can enhance the output's alignment with their preferences, albeit at the expense of an increased communication cost. In short, users face a trade-off between AI *output fidelity*[1] and *communication cost*.[2]

Different users may respond to this trade-off differently, but those who value productivity more than fidelity will rely on AI more and willingly let go of their own preferences. We are interested in the potential societal consequences of these choices. First, working with AI may be more beneficial for some users than others: in the Copilot example, users who prefer the default output would not even need to communicate with the AI to have high fidelity to their preferences. Second, as users do not share complete preferences with the AI and let it "choose" for them, the produced content may be, on average, *homogenized* towards the AI's default choices. For example, ChatGPT has been trained with reinforcement learning from human feedback (RLHF) (Kinsella, 2023) to have a specific tone and language. If

[1]Notice that *output fidelity* does not measure the "quality" of the output; instead, it measures how much the output deviates from a user's personal preference. In particular, we focus on scenarios where users already know how to complete the task without AI, and the output is always correct (e.g., a software engineer's code does the job). However, there are many ways to achieve the same outcome (e.g., the engineer's coding style).

[2]In fact, the importance and the associated costs of communicating with AIs have given rise to a new profession called prompt engineering (Mok, 2023), and spurred the creation of novel marketplaces like PromptBase (URL: https://promptbase.com/).

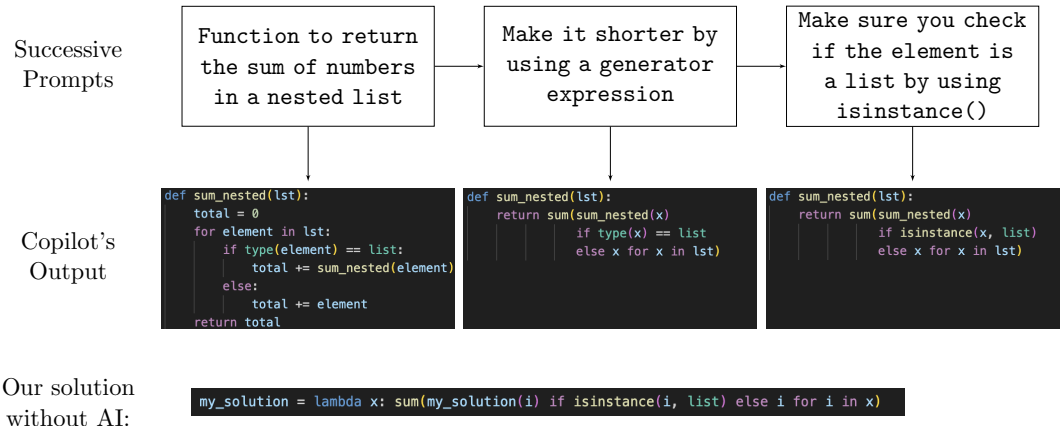


Figure 2.1: The incremental information provided to align GitHub Copilot's Python code output with our preference. While the initial output in Attempt 1 is functional, it significantly differs from our solution without AI. Bridging the gap requires several iterations.

students use ChatGPT's help for their homework, their writing style may be influenced by ChatGPT's. More generally, AIs are built by a few but used by many, and there is a risk any AI bias could turn into a *societal bias*. The AI training process may involve censoring (e.g., the choice of the dataset) and human input (e.g., RLHF), which could intentionally or unintentionally lead to bias. For example, some studies discuss ChatGPT's inclination towards left-leaning political stances (Hartmann et al., 2023; Motoki et al., 2023; Rozado, 2023). All in all, because of the benefits of increased productivity and the balance between output fidelity and communication costs, users could willingly produce less diverse content that is vulnerable to potential AI biases.

We propose a Bayesian model to study the societal consequences of human-AI interactions. For a given task, rational users can exchange information with the AI to align its output with their heterogeneous preferences. The AI has a knowledge of the distribution of preferences in the population and uses a Bayesian update to create the optimal output with maximal expected fidelity given the information shared by the user. Users choose the amount of information they share to maximize their utility, balancing the cost of communication with the fidelity of the output. In this setting, we aim to formalize and evaluate the

societal risks of homogenization and AI bias and how they could be mitigated.

When solving for each user's optimal decision, we find that their use of AI depends on how "unique" they are. Users with more common preferences simply accept the default output, avoiding any communication costs at the expense of a small fidelity mismatch. In contrast, users with more unique preferences share information with the AI to reduce fidelity errors, albeit with higher communication costs. And the most unique users do not benefit from the AI and simply perform the task themselves. Interestingly, we find that for less common users, their fidelity improves with the uniqueness of their preferences. In other words, when we compare two users with relatively unique preferences (neither the most common nor the most unique), the user with the most unique preferences between the two will experience better fidelity.

To establish the homogenization effect, we prove that any output resulting from human-AI interactions is less unique than what a user would have done without AI. This is confirmed at the population level, where the AI-generated output distribution has a lower variance than the users' preference distribution. This phenomenon is exacerbated when AI-generated content is used to train the next generation of AI: we show that the users' rational decisions and the AI's training process can mutually reinforce each other, leading to a homogenization "death spiral." However, we also demonstrate that facilitating human-AI interaction by offering improved means for users to express their preferences can serve as a valuable tool in mitigating this effect and preserving population diversity. For example, Open AI has experimented with custom instructions (OpenAI, 2023a) and voiced-based interactions (OpenAI, 2023c), while Jina AI offers tools that optimize prompts.[3]

We also study the effects of AI bias, identifying who benefits or loses when using an AI model that does not accurately reflect the population preference distribution. At the population level, the censoring type of bias (e.g., biasing against the more unique preferences)

[3]E.g., automatic prompt optimization such as Prompt Perfect (URL: https://promptperfect.jina.ai/).

negatively impacts the population utility as whole, especially users with uncommon preferences who rely on AI interactivity the most. This may seem counter-intuitive as we might assume that the majority with common preferences would benefit from censorship. Yet, our findings reveal that the benefits for this majority are marginal, while the harm to the minority with unique preferences is substantial, leading to an overall loss in the population utility. On the other hand, directional biases (e.g., a slightly left-leaning AI) are not as harmful in terms of utility, but any directional bias will influence the users' chosen output, leading to a societal bias. On the positive side, the user interactions with the AI partially counter the effects of AI bias, highlighting the need to consider human decisions to fully understand the impact of generative AI.

We show that tasks that are either hard to do without AI (e.g., image generation) or for which speed is particularly important (e.g., grammar correction) are especially sensitive to the risks of homogenization and bias. However, our research demonstrates that creating models that facilitate human-AI interactions can significantly limit these risks and preserve the population preference diversity.

2.2 Literature review

Related studies on the issues of homogenization and bias. A few studies have a focus related to the homogenization issue (e.g., Anderson et al. (2024); Bommasani et al. (2022); Chaney et al. (2018); Doshi and Hauser (2024); Padmakumar and He (2024); Saatci and Wilson (2017); Shumailov et al. (2023)). Recent empirical research indicates that generative AI may reduce the diversity of outputs, which aligns with our findings (Anderson et al., 2024; Doshi and Hauser, 2024; Padmakumar and He, 2024). For example, Doshi and Hauser (2024) found that while generative AI can improve the quality and enjoyment of written articles, it also makes the stories more similar to each other than those written solely by humans. Some other studies have examined how the training process of an AI may reduce

the diversity of AI-generated content (Bommasani et al., 2022; Shumailov et al., 2023). Shumailov et al. (2023) observed that the tails of the original content distribution disappear when AIs are successively trained from AI-generated content (they call it model collapse). Also, Bommasani et al. (2022) demonstrate that algorithmic systems built on the same data or models tend to homogenize outcomes. Moreover, before the launch of ChatGPT, a similar homogenization issue has also been discussed in the literature of recommendation systems (Chaney et al., 2018). By using a simulation, Chaney et al. (2018) show that a feedback loop, where a recommendation system is trained on data from users already exposed to AI recommendations, may homogenize user behavior.

On the other hand, the issue of bias in generative AI has also been shown (Hartmann et al., 2023; Motoki et al., 2023; Rozado, 2023), with empirical evidence of its impact on cognitive processes (Bhat et al., 2023; Jakesch et al., 2023). For example, Rozado (2023) implemented 15 different political orientation tests to ChatGPT. The author found that ChatGPT's answers manifested a preference for left-leaning opinions in 14 of the 15 tests. Bhat et al. (2023) discovered that people may incorporate AI suggestions into their writing, even when they disagree with the suggestions overall. Similarly, Jakesch et al. (2023) showed that biased language models could influence the opinions expressed in people's writing and shift their viewpoints.

In contrast to these studies, our research examines the causes of homogenization and bias from a human-centric perspective by using a modeling approach. Specifically, we employ a Bayesian model to explore how users' rational decision-making when interacting with AIs affects these issues. Our findings underscore the importance of enhancing AI usability and encouraging users to share more information to address these societal challenges. To the best of our knowledge, our paper is the first modeling study that employs such a human-centric perspective to understand the societal consequences of generative AIs.

Related studies on human-AI interactions. Our paper is also related to some recent modeling studies about human-AI interaction in operations management (e.g., Agrawal et al. (2018); Bastani et al. (2022); Boyacı et al. (2023); Chen et al. (2022); Dai and Singh (2023); de Véricourt and Gurkan (2023); Ibrahim et al. (2021); Mclaughlin and Spiess (2023)). Essentially, their primary focus lies in examining the potential impact of the coexistence of humans and an AI on performance, such as accuracy, and exploring how the predictive performance can be enhanced or hindered compared to decisions made solely by humans or AI. For example, de Véricourt and Gurkan (2023) consider the human-AI interactions in which a human agent supervises an AI to make some high-stakes decisions. They show that the agent may be subject to a verification bias and hesitates forever whether the AI performs better than the agent because the agent can overrule the AI before observing the correctness of the AI's predictions. Ibrahim et al. (2021) build a stylized model to analyze how human judgments can improve AI predictions. In the paper of Boyacı et al. (2023), the authors consider a situation in which a human agent has to spend a cognitive cost collecting information in a decision process, whereas an AI can provide him with some additional information without cognitive cost. They show that the AI input can improve the overall accuracy of human decisions but may incur a higher propensity for certain types of errors. These papers primarily focus on decision-making when human and AI options exist separately. In contrast, our paper considers a more interactive setting, where users can provide a generative AI with more information to improve the AI's outputs but have to spend a communication cost.

Related studies on generative AIs. With the popularity of ChatGPT, many scholars have engaged in research on its impact on people's lives and in their respective fields, such as labor markets (Eloundou et al., 2023), marketing (Brand et al., 2023), healthcare (Sallam, 2023), and so on. Most of the research uses empirical analysis to investigate whether generative AI, represented by ChatGPT, can truly bring us more benefits and conveniences.

For instance, Noy and Zhang (2023) show that ChatGPT can substantially improve productivity in mid-level professional writing tasks. Binz and Schulz (2023) tested GPT-3 with some experiments from the cognitive psychology literature. They find that GPT-3 can solve many of those tasks well and even sometimes outperform humans' performance. Our study approaches this question from a different angle. Through a modeling method, we attempt to foreshadow how our lives may change under the widespread application of generative AIs due to people's rational decision-making when interacting with AIs.

We assume that the output of AIs depends on the information provided by users. In fact, many empirical studies have observed that AIs are quite sensitive to users' inputs (Binz and Schulz, 2023; Brand et al., 2023; Liu et al., 2023). For example, Brand et al. (2023), who adopted ChatGPT to conduct marketing research, found that GPT is sensitive to the phrasing of queries in their empirical work. When querying GPT with a list of options, they found that GPT is more likely to choose the first option. Denny et al. (2023) also indicated that sending proper prompts is critical for the performance of Copilot.

Related studies on the modeling approach. The way we model the human-AI interaction shares similarities with the frameworks of information design (Kamenica and Gentzkow, 2011), costly persuasion (Gentzkow and Kamenica, 2014), the theory of rational inattention (Sims, 2003), as well as the interpretation of LLMs with Bayesian inference (Wei et al., 2021; Xie et al., 2022). The user's decision is modeled similarly to an information design process (Alizamir et al., 2020; de Véricourt et al., 2021). The sender (i.e., the user) sends a signal to the receiver (i.e., the AI) to inform the receiver about a true state (i.e., the user's preference). The utility of the sender is determined by the receiver's decision (i.e., the AI's output). Additionally, we employ the framework of costly persuasion (Gentzkow and Kamenica, 2014) and the theory of rational inattention (Sims, 2003) to model the user's communication cost when sending the signal. In particular, we follow the standard way in the literature to model the cost of information as the expected reduction in entropy. This

assumption can also be found in other modeling papers, such as the cognitive cost defined in Boyacı et al. (2023). Note that we assume the reduction in entropy is relative to the population distribution of users' preferences (see Section 2.3) instead of AI's prior (defined in Section 2.3). As Gentzkow and Kamenica (2014) suggested, the reduction in entropy can be defined relative to any proper fixed reference belief. So we use the population distribution of users' preferences as the fixed reference belief to indicate that the communication cost is independent of AI's prior but relevant to the difficulty of distinguishing a user's preference from the others. Furthermore, we model the AI's behavior as a Bayesian inference (Wei et al., 2021; Xie et al., 2022). For instance, Xie et al. (2022) interpret that the in-context learning of an LLM can be viewed as an implicit Bayesian inference. The prior of the LLM is formulated during training. Conditional on a prompt, the LLM characterizes a posterior distribution to make an output.

2.3 Model Setup

We develop a Bayesian model to represent the process of working with generative AI on a given task. Users have preferences on how to complete the task, and the AI knows the population's distribution of these preferences (through its training). Each user can also interact with the AI to share information about her specific preferences. This interaction will help the AI produce an output closer to what a user would have done without AI, leading to a better output *fidelity*. However, sharing information requires effort, which entails a *communication cost*. Users must choose how much information they share to balance the benefits of fidelity and the cost of communication.

User preferences and Fidelity We use $\theta \in \mathbb{R}$ to denote a user's specific *preference*, and we assume that θ is normally distributed in the population: $\theta \sim N(\mu_\theta, \sigma_\theta^2)$. Here μ_θ represents the average population preference and σ_θ the diversity of the population's

preferences. In practice, the users' preferences should be represented by a high-dimensional space, but we will interpret θ as a specific *feature* of a user's preferences, as illustrated in the following example.

Example 1 (News Article)**.** A journalist would like to write an article about a piece of breaking news and wants to use ChatGPT to write faster. θ measures the political orientation of the article this journalist would have written without AI. If $\theta > \mu_\theta$, the journalist is more right-leaning than the average journalist. When using AI, the journalist may be able to write faster but may not meet her exact political orientation θ (low fidelity).

We will refer to a "user θ" to describe a user with preference θ, and to an "output θ_A" to refer to an AI's output matching some preference θ_A. We define the output's *fidelity loss* as $(\theta - \theta_A)^2$, and we interpret it as a loss in utility for a user θ receiving an output θ_A (users prefer an output matching their preference). For example, if $\theta_A = \theta - 1$, the AI of Example 1 outputs a more left-leaning article than the journalist's preferred political orientation, and the fidelity loss is 1.

AI Bayesian Inference We model the interaction between a user and AI as an exchange of information about θ. The AI has a prior belief of the population distribution of θ. This belief corresponds to a normal distribution $N(\mu_A, \sigma_A^2)$ with density $\pi_A(\cdot)$. To capture that the AI has been trained on a representative dataset, we assume that the AI's prior is exactly the population distribution, $\mu_A = \mu_\theta$ and $\sigma_A = \sigma_\theta$ (this assumption is relaxed in Section 2.6 to study the effects of a biased AI). We model the exchange of information between a user θ and the AI as with normal distributions: the user shares a noisy signal $q = \theta + \epsilon_q$ where $\epsilon_q \sim N(0, \sigma_q^2)$, and the AI refines its belief using Bayes' rule: $\theta|q \sim \pi_A(\cdot|q)$. It then returns the optimal output with the maximum expected fidelity:

$$\theta_A \triangleq \arg\min_{\hat{\theta}} E[(\hat{\theta} - \theta)^2|q] = E[\theta|q] = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_q^2} \cdot q + \frac{\sigma_q^2}{\sigma_A^2 + \sigma_q^2} \cdot \mu_A. \tag{2.1}$$

Note that θ_A is a weighted average between q and the prior mean (Berger, 1985), which is a random variable since q is noisy. Additionally, a lower value of σ_q corresponds to more information shared with the AI: if $\sigma_q = 0$, then the user θ shares her exact preference, and the AI returns $\theta_A = \theta$. In the limit $\sigma_q \to +\infty$, the signal is uninformative, and the AI outputs the mean of its prior, $\theta_A = \mu_A$.

The *fidelity error* of a user θ given σ_q is the expected output fidelity, denoted by:

$$e(\theta, \sigma_q) \triangleq E[(\theta_A - \theta)^2 | \theta]. \tag{2.2}$$

We can decompose it into two terms,

$$e(\theta, \sigma_q) = Var(\theta_A|\theta) + [E(\theta_A|\theta) - \theta]^2.$$

For a user θ, the first term corresponds to the variability in the AI's output that stems from the information exchange, while the second term is the impact of the bias in the AI response, which will be a focus of the paper. In the context of Example 1, the bias is high if the AI-written article consistently leans more to the left than the journalist orientation.

Information and Communication Cost Given an exchange of information parametrized by σ_q, we measure the "communication cost" of the user to share this information with the AI. Following standard assumptions in the rational inattention (Sims, 2003) and costly persuasion (Gentzkow and Kamenica, 2014) literature, we assume the communication cost to be proportional to the expected reduction in the AI uncertainty of θ relative to the population distribution of θ given σ_q:

$$\lambda I(\sigma_q) \triangleq \lambda[H(\theta) - E[H(\theta|q)]] = \lambda \left[\ln(\sigma_\theta \sqrt{2\pi e}) - \ln \left(\sqrt{\frac{\sigma_\theta^2 \sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}} \sqrt{2\pi e} \right) \right] = -\frac{\lambda}{2} \ln \left(\frac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2} \right),$$

where $\lambda > 0$ is the marginal cost of communication, $I(\sigma_q)$ is the mutual information, and $H(\cdot)$ denotes the differential entropy. Intuitively, $I(\sigma_q)$ corresponds to the "amount of information" the user shares about her preference θ. Note that sharing the exact value of

θ ($\sigma_q = 0$) requires an *infinite* amount of information $I = +\infty$. Conversely, providing an uninformative signal about θ ($\sigma_q \to +\infty$) requires no information $I = 0$. To interpret this situation, remember that our model assumes that the AI knows the task description and that the human-AI interaction is about sharing user preferences. Without preference information, the AI uses its knowledge of the preference distribution to return a "default" output for the task, $\mu_A = \mu_\theta$ (e.g., the first answer of ChatGPT) . In Example 1, ChatGPT would write an article that expresses an "average" political orientation that does not necessarily reflect the journalist's views. And, in the GitHub Copilot coding example of Figure 2.1, the initial Copilot's output is a default function that does not consider the user's specific preference. In both cases, the AI requires more information to deliver better fidelity.

User's decision Each user θ chooses the information I they share with the AI (parametrized by σ_q) to minimize their *utility loss* l given by the sum of the fidelity error and the communication cost

$$l(\theta, \sigma_q) \triangleq e(\theta, \sigma_q) + \lambda I(\sigma_q). \tag{2.3}$$

That is, a user θ chooses an optimal $\sigma_q^\star(\theta)$ that minimizes her utility loss,

$$\sigma_q^\star(\theta) \triangleq \arg\min_{\sigma_q \geq 0} l(\theta, \sigma_q). \tag{2.4}$$

Importantly, λ controls the tradeoff between fidelity error and communication cost, and we will refer to it as the *cost of human-AI interactions*. A task has a low λ if it is particularly easy to interact with the AI and share preferences (e.g., a chat interface like ChatGPT or voice-based interactions (OpenAI, 2023c)) and/or if users care a lot about fidelity. λ will be high if users care more about minimizing effort than matching their specific preferences. Because the task of Example 1 is about breaking news, the journalist may be in a hurry and have a high λ.

Choosing to work with AI If the cost of human-AI interaction is high and fidelity is important, a user might be better off not using the AI and doing the work herself. In this

case, the output would have no fidelity error by definition. However, manual work takes time, which we model as a fixed utility cost $\Gamma > 0$ that depends on the task but is the same for everyone.

If Γ is smaller than the expected utility loss $l(\theta, \sigma_q^\star)$, then a user θ will not use the AI. We define the output $\theta^\star$ chosen by a user θ and the corresponding expected utility loss l^* as:

$$\theta^\star \triangleq \begin{cases} \theta_A|(\theta, \sigma_q^\star) & \text{if } l(\theta, \sigma_q^\star) \leq \Gamma \\ \theta & \text{otherwise} \end{cases}, \qquad l^* \triangleq \min(l(\theta, \sigma_q^\star), \Gamma). \tag{2.5}$$

2.4 Human-AI Interactions and Homogenization

A consequence of our model is that different users may interact with the AI differently, sharing varying amounts of information about their preferences or even choosing not to use the AI. We first describe these individual-level choices and then study their implied societal consequences and how to mitigate them.

2.4.1 Individual Level: Heterogeneous Use of AI

Analyzing the optimal decision of each user θ requires solving Problem (2.4), and the results are presented in Proposition 1. Users' choices depend on their *uniqueness*, the distance of their preference θ to the population mean μ_θ, $d(\theta) \triangleq |\theta - \mu_\theta|$. We note that the derivation of $\sigma_q^\star(\theta)$, presented in Appendix A.1, is non-trivial as Equation (2.3) is neither concave nor convex.[4]

Proposition 1. Optimal user strategies solving Equation (2.5) have the following characteristics:

[4]We present the proofs for all the statements in Appendix A.2.

1. More unique users have a higher utility loss: $l^\star$ increases[5] in $d(\theta)$.

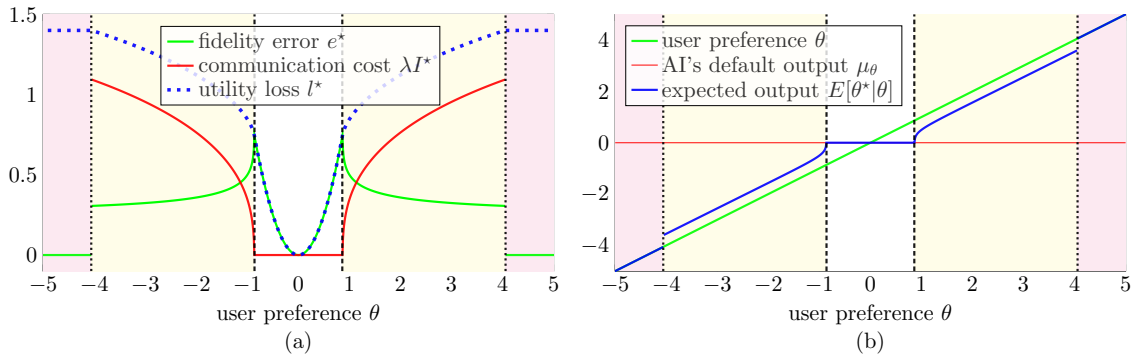
2. More unique users interact more with the AI (if they choose to use it): $\lambda I^\star$ increases in $d(\theta)$.

3. Users use the AI if they are below a uniqueness threshold $\tau_a > 0$: $d(\theta) \leq \tau_a \iff l(\theta, \sigma_q) \leq \Gamma$.

4. Users that use AI are characterized by another uniqueness threshold τ_d such that:

   (a) If $d(\theta) \leq \tau_d$, users choose the default AI output ($I^\star = 0$) and their fidelity error $e(\theta, \sigma_q^\star)$ increases with their uniqueness $d(\theta)$.

   (b) If $d(\theta) > \tau_d$, users interact with AI ($I^\star > 0$) and their fidelity error *decreases* with their uniqueness.



(a) (b)

Figure 2.2: The black dashed vertical lines are at $d(\theta) = \tau_d$, and the black dotted vertical lines are at $d(\theta) = \tau_a$. The white region indicates the users who choose the default output; the yellow region indicates those who send information to the AI; the red region indicates those not using AI. We use $\mu_\theta = 0, \sigma_\theta = 1, \lambda = 1, \Gamma = 1.4$.

[5] All references to "increasing" or "decreasing" functions are meant in a weak sense (i.e., "non-decreasing").

The main takeaway from Proposition 1 is that users with more "common" preferences have a utility advantage (Item 1) and choose to interact less with the AI (Item 2). The fundamental driver of this, crucial throughout the paper, is that more common users can have a small fidelity error with limited shared information. There are three types of users: users who use AI but do not share information, those who share information, and users who do not use AI. The most common users, with $d(\theta) \leq \tau_d$ as described in Item 4a, accept the default output of the AI, zero communication cost but rapidly increasing fidelity error as these users become more unique, and the default output $\theta_A = \mu_\theta$ becomes worse (see the region in the center of Figure 2.2. Users with $\theta > \tau_d$ then choose to interact with the AI (Item 4b), which will reduce their fidelity error at the expense of communication cost (Item 2). Indeed, as illustrated in Figure 2.2 (a), while the fidelity error (green curve) dominates the utility loss of users with common preferences, more unique users prefer to pay an increasing communication cost (red curve) that dominates a decreasing fidelity error. Interacting with the AI eventually reaches such high communication costs for the most unique users ($d(\theta) > \tau_a$) that the no-AI option becomes preferable (Item 3) as shown in the red area of Figure 2.2 (a).

Many users have a positive fidelity error, so the AI's output may not always align perfectly with a user's preference. The next proposition shows that this output is misaligned in a specific way: on average, a user's chosen output $\theta^\star$ tends to revert toward the population's mean preference.

Proposition 2. The expected chosen output $E[\theta^\star|\theta]$ of any user θ is closer to the population's mean than their actual preference: $|\mathbb{E}[\theta^\star|\theta] - \mu_\theta| \leq |\theta - \mu_\theta|$. Moreover, the inequality is strict for almost all users that use the AI: when $d(\theta) < \tau_a$ and $\theta \neq \mu_\theta$.

We illustrate this result in Figure 2.2 (b). The most common users (with $d(\theta) \leq \tau_d$) provide an uninformative signal to the AI ($I^\star = 0$) and accept the AI's default output $\theta_A = \mu_\theta$, which is a direct revert to the mean. As users become more unique, they interact with the AI ($I^\star > 0$), which mitigates the mean reversion in the AI's output. However,

it doesn't completely vanish due to the high communication cost. The mean reversion disappears only for those very unique users who choose to work themselves and not to use the AI. In Example 1, a journalist with $\theta > \mu_\theta$ would write a (slightly) more left-leaning article than her preference. As discussed in the next section, this can be an issue at the population level.

2.4.2 Societal Level: Homogenization

In a world without AI, the distribution of people's output would exactly match the distribution of their preference $\theta \sim N(\mu_\theta, \sigma_\theta^2)$. However, with AI, the output is $\theta^\star$, which does not have the same distribution, as we saw that users of AI tend to choose an output closer to the mean μ_θ. At the population level, this leads to *homogenization*, where the output distribution has a lower variance than the distribution of preferences.

Theorem 1. *When everyone uses AI ($\Gamma \to +\infty$), the variance of the population output is lower than the variance of the population preferences, $Var(\theta^\star) < Var(\theta)$, and strictly decreases in the cost of human-AI interactions λ. In the general case ($\Gamma < +\infty$), $\lim_{\lambda \to 0} Var(\theta^\star) = Var(\theta)$ and $\lim_{\lambda \to +\infty} Var(\theta^\star) < Var(\theta)$.*

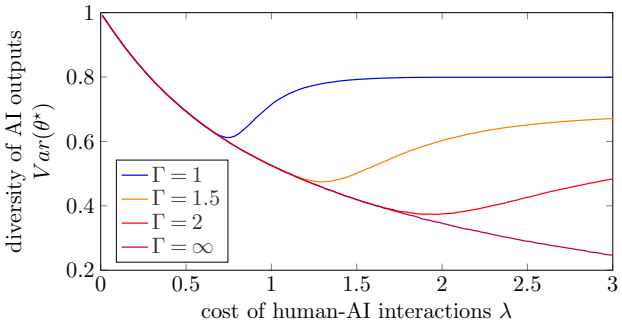


Figure 2.3: We use $\mu_\theta = 0$, $\sigma_\theta = 1$.

Theorem 1 formalizes the risk of homogenization and points to its solution. When everyone uses AI ($\Gamma \to +\infty$), reducing the cost of human-AI interactions λ encourages users

to interact more with the AI and share their specific preferences more accurately, limiting homogenization and helping to preserve the population's diversity. When $\lambda \to 0$, there is no more homogenization as users can share their precise preference for free. The case $\Gamma < +\infty$ is more involved, as some users renounce the AI when the cost of human-AI interactions is high, partially improving the chosen output's diversity. We illustrate it in Figure 2.3 and present a more in-depth analysis in Appendix A.5. An interesting special case is when $\Gamma < +\infty$ and $\lambda \to +\infty$. Only two types of users remain: those who complete the task themselves and those who accept the default AI output, leading to homogenization on average. In all cases, Theorem 1 underscores that enhancing the interactivity of AI tools (e.g., through better interfaces, multi-modal inputs, or real-time feedback mechanisms) to achieve a sufficiently low λ is an effective strategy to encourage users toward higher fidelity, reduce homogenization, and ultimately, preserve population preference diversity.

2.5 AI-generated content and the "Death Spiral" of Homogenization

While homogenization can easily be perceived as a negative societal outcome, we argue it may also have long-term consequences. As more and more content becomes AI-generated, this content could be used to train the next generation of AI. Because of the homogenization issue, this would lead to an incorrect AI distribution of human preference (the AI's prior). The next AI generation would be even more likely to return homogenized outputs, resulting in a "death spiral" of homogenization, a dreadful outcome for human preference diversity, where the diversity of outputs is diminishing over time. We study this phenomenon within our model, considering a *self-training loop* where the population's output distribution at time t becomes the AI prior at time $t + 1$ (as illustrated in Figure 2.4).

This model is not tractable, as our analysis does not apply when the AI has a non-normal prior, which happens after the first iteration of the self-training loop. Therefore, we

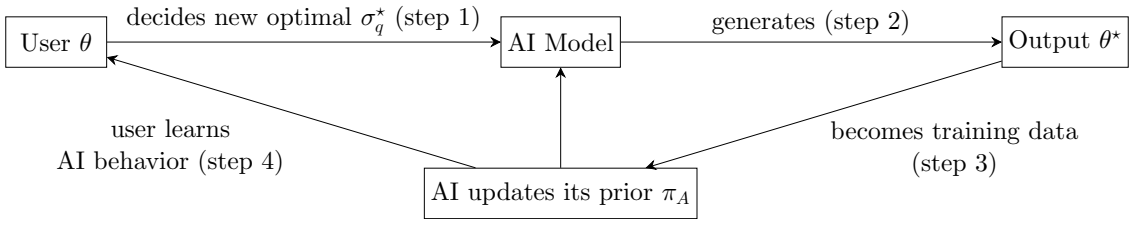


Figure 2.4: Steps in each iteration of the self-training loop.

further simplify our model to understand the effects within a self-training loop. We then use simulations to verify the theoretical results from the simplified model and conduct robustness tests in more complex scenarios.

2.5.1 A simplified model

Since the difficulty arises from non-normal priors after the first iteration, we assume that the user preference θ follows a three-point distribution with support $\Theta \triangleq \{-v, 0, v\}$ and a probability mass at zero p_0:

$$\pi_\theta(\theta) = \begin{cases} (1-p_0)/2 & \text{if } \theta = -v \\ p_0 & \text{if } \theta = 0 \\ (1-p_0)/2 & \text{if } \theta = v \end{cases}$$

Let $\pi_A^t(\theta)$ denote the AI prior at time t and $\pi_A^t(\theta|q, \sigma_q)$ denote the posterior after receiving a signal $q \sim \theta + \epsilon_q$ where $\epsilon_q \sim N(0, \sigma_q)$. In line with the original model setup, the AI output given q at time t maximizes the expected fidelity:

$$\theta_A^t \triangleq \arg\min_{\hat{\theta}\in\Theta} E[(\hat{\theta}-\theta)^2|q] = \arg\min_{\hat{\theta}\in\Theta} \sum_{\theta\in\Theta} (\hat{\theta}-\theta)^2 \pi^t(\theta|q,\sigma_q)$$

The user's decision is also as defined in Section 2.3. That is, a user θ needs to find $\sigma_q^{\star t}$ that solves

$$\min_{\sigma_q} l^t(\theta, \sigma_q) = e^t(\theta, \sigma_q) + \lambda I(\theta, \sigma_q)$$

where $e^t(\theta, \sigma_q) = E[(\theta_A^t - \theta)^2|\theta]$ is the fidelity error and $\lambda I(\theta, \sigma_q) = \lambda[H(\theta) - E[H(\theta|q)]]$ is the communication cost. And a user θ can still choose to work without the AI if the utility

loss of using AI is too high. As defined in Section 2.3, the output $\theta^{\star t}$ is:

$$\theta^{\star t} = \begin{cases} \theta_A^t | (\theta, \sigma_q^\star) & \text{if } l^t(\theta, \sigma_q^\star) \leq \Gamma \\ \theta & \text{otherwise} \end{cases}$$

Importantly, in a self-training loop, the AI outputs are reused to train the next generation of AI, so this means that the AI prior at time $t+1$ becomes the unconditional distribution of $\theta^{\star t}$:

$$\pi_A^{t+1}(\theta) \triangleq \begin{cases} \mathbb{P}\left(\theta^{\star t} = -v\right) & \text{if } \theta = -v \\ \mathbb{P}\left(\theta^{\star t} = 0\right) & \text{if } \theta = 0 \\ \mathbb{P}\left(\theta^{\star t} = v\right) & \text{if } \theta = v \end{cases}$$

Note that the initial AI prior is still assumed to be the same with the population distribution of θ (i.e., $\pi_A^0(\theta) = \pi_\theta(\theta)$). We then define the phenomenon of the homogenization death spiral as follows.

Definition 1 (Homogenization Death Spiral)**.** The homogenization death spiral is a phenomenon where the variance of outputs, $Var(\theta^{\star t})$, is monotonically decreasing in time t.

This model simplifies the original model in a self-training loop but is still able to maintain the key properties. Users are still facing a trade-off between fidelity error and communication cost, defined as before. Users' preferences remain heterogenous: some preferences are more unique (i.e., $\theta = -v$ and $\theta = v$), while the others are more common ($\theta = 0$). We refer to $\theta = 0$ as the common users and to $\theta = -v$ or $\theta = v$ as the unique users. The only difference is that the population distribution of θ and the AI prior are constrained to a discrete support with three points. This simplification enables us to further analyze the effects of a self-training loop and how a homogenization death spiral emerges.

2.5.2 Factors affecting the homogenization death spiral

With the simplified model, we are able to comprehend the driving forces behind a homogenization death spiral. As a preliminary result, the following lemma illustrates the behavior

of the common users and the symmetry of the AI prior.

Lemma 1. It is optimal for the common users to accept the default output, and the AI prior remains symmetric. That is, $\forall t,\ \sigma_q^{\star t}(0) = \infty$ and $\pi_A^t(-v) = \pi_A^t(v)$.

Lemma 1 is intuitive because the common users can achieve zero utility loss by accepting the default output without making any effort. Given σ_q, the unique user's utility loss is the same, whether $\theta = -v$ or $\theta = v$, as long as the AI prior at time t is symmetric, leading to a symmetric AI prior in the next iteration. Building on Lemma 1, we can prove the following corollary.

Corollary 1. $\forall t,\ Var(\theta^{\star t}) \leq Var(\theta)$, and $Var(\theta^{\star t}) = Var(\theta)$ if and only if $\sigma_q^{\star t}(-v) = \sigma_q^{\star t}(v) = 0$.

We can view Corollary 1 as an analogy to Theorem 1 in the simplified model. It demonstrates that the diversity of outputs is reduced because users cannot fully exert effort to share information about their preferences.

With the above foundations, let us now focus on a single iteration with any symmetric AI prior $\pi_A^t(\theta)$. This analysis will help us understand how the AI prior at time $t+1$ depends on the previous iteration at time t. The following proposition illustrates how the variables at time t may affect the variance of outputs at time $t+1$. Note that Proposition 3 assumes $\Gamma = \infty$ and does not consider the user's optimization problem to isolate the effect of variables.

Proposition 3. Suppose $\Gamma = \infty$ and $\pi_A^t(-v) = \pi_A^t(v)$. Holding $\sigma_q^t(-v) = \sigma_q^t(v) = \sigma_q$ for some σ_q, we have:

1. $Var(\theta_A^{t+1})$ monotonically increases in $Var(\theta_A^t)$.
2. $Var(\theta_A^{t+1})$ monotonically decreases in σ_q.

The first result in Proposition 3 indicates that an increase or decrease in the variance of outputs has a lasting impact, influencing the variances of outputs in subsequent periods

in the same direction. Intuitively, if the AI focuses predominantly on the majority and its prior becomes more concentrated around the average, it becomes more difficult for unique users to reduce fidelity error. Consequently, the AI is more likely to generate outputs close to the average, further concentrating the distribution of outputs around the average. On the other hand, the second result in Proposition 3 suggests that making efforts to share more information acts as a counterforce against homogenization, increasing the variance of outputs. As previously illustrated in Section 2.4.2, sharing more information effectively preserves the diversity of outputs and mitigates homogenization in the first period. Proposition 3 demonstrates that this effect of information sharing is consistent across all periods in a self-training loop. Essentially, this proposition highlights the long-term impact of users' efforts in maintaining output diversity. If users keep σ_q constant and do not react to homogenized outputs in the current iteration, this homogenization issue will propagate through all future iterations, reducing output diversity within each period.

Nonetheless, Proposition 3 does not explain how the variance of outputs may change period by period. The following theorem illustrates the major forces that lead to or prevent the homogenization death spiral.

Theorem 2. *When everyone uses AI, the homogenization death spiral exists if users cannot share more information than the last iteration. That is, when* $\Gamma = \infty$, $\forall t,\ Var(\theta^{\star(t+1)}) \leq Var(\theta^{\star t})$ *if* $\sigma_q^{t+1}(\theta) \geq \sigma_q^t(\theta)$ *for any* θ. *In contrast, there exists a case where the homogenization death spiral does not exist if one of the following conditions is true:*

1. Γ *is small enough.*
2. $\Gamma = \infty$ *but the users with* $\theta \neq 0$ *share sufficiently more information than the previous iteration.*

The first part of Theorem 2 highlights the importance of sharing sufficient information to maintain output diversity in a self-training loop. If users cannot share more information than in the last iteration, the diversity of outputs will be lower than in the last iteration.

As demonstrated in Proposition 3, since the AI always tends to cater to the majority, the distribution of generated outputs will be more concentrated on the average than the AI prior if users cannot share extra information. This further makes it more difficult for users to reduce their fidelity errors in subsequent iterations, creating a snowball effect that monotonically reduces output diversity. However, the second part of Theorem 2 identifies factors that counteract this reduction in diversity. First, when Γ is finite, there are scenarios where users choose to use the AI at time t but opt to work without the AI at time $t+1$. Manual work can restore output diversity, leading to a higher variance than the previous iteration and breaking the homogenization death spiral. Additionally, even if all the users continue to use the AI, there are cases where users are willing to exert sufficient effort and share enough information if the cost of human-AI interaction is low enough. Sharing sufficient information can also preserve output diversity and disrupt the homogenization death spiral. Overall, these counterforces interact to each other, potentially resulting in varying levels of output variance across iterations.

Theorem 2 demonstrates that manual work and sharing sufficient information are two effective strategies for breaking the homogenization death spiral. Otherwise, the homogenization death spiral persists, and the diversity of outputs will continue to diminish over time. From a social planner's perspective, this implies the importance of encouraging individual contributions without AI and facilitating efficient human-AI interactions to prevent the homogenization death spiral.

The limitation of Theorem 2 is that it does not precisely indicate how the diversity of outputs changes over time due to the complexity of solving the users' optimization problems, even in the simplified model. Nonetheless, we further investigate the existence of the homogenization death spiral in more complex settings by using numerical methods as robustness tests, which are presented in the following section.

2.5.3 Robustness tests

In the numerical experiments, we first discretize the continuous population distribution of θ and the continuous distribution of q by the Lloyd-Max algorithm (Gallager et al., 2008) to achieve computational tractability. Starting with the initial prior $\pi_A^0(\theta) = \pi_\theta(\theta)$, we numerically compute the posterior $\pi_A^t\theta|q, \sigma_q$ for given q and σ_q. This enables us to determine $\theta_A^t|q, \sigma_q$. We then derive the fidelity error and the communication cost, as defined in Section 2.3 for any σ_q. Subsequently, we numerically find the optimal $\sigma_q^\star$ for each θ at time t and calculate $\mathbb{P}(\theta^{\star t} = \theta)$. At the beginning of the next iteration, the AI prior is updated to $\pi_A^{t+1}(\theta) = \mathbb{P}(\theta^{\star t} = \theta)$. These steps are repeated until a specified number of iterations is reached. A detailed description can be found in Appendix A.3.

The original model Let us first revisit our original model where the population distribution of θ and the AI prior are normally distributed. As shown in Figure 2.5, if everyone uses the AI (i.e., $\Gamma = +\infty$), the variance of outputs the variance of outputs decreases over time. This decrease is most pronounced during the first iteration when users initially begin utilizing the AI. Then, there is a slight recovery in variance as users share more information than they did in the first iteration. However, this is short-lived, and the "death spiral" persists, leading to a continuous and monotonic decrease in output variance. This occurs because users' efforts are insufficient to enhance the diversity of outputs, causing the distribution of generated outputs to converge increasingly toward the mean. This means that as the AI's prior becomes increasingly erroneous, the communication cost necessary to reduce the fidelity error becomes unmanageable, and more and more users start to accept the default output until we converge to a more and more homogenized world.

Nonetheless, we can also observe that the homogenization death spiral can be mitigated when the cost of human-AI interaction, λ, is small or the manual cost, Γ, is low. As illustrated in Figure 2.5 (a), when everyone uses the AI, a lower λ results in a higher curve,

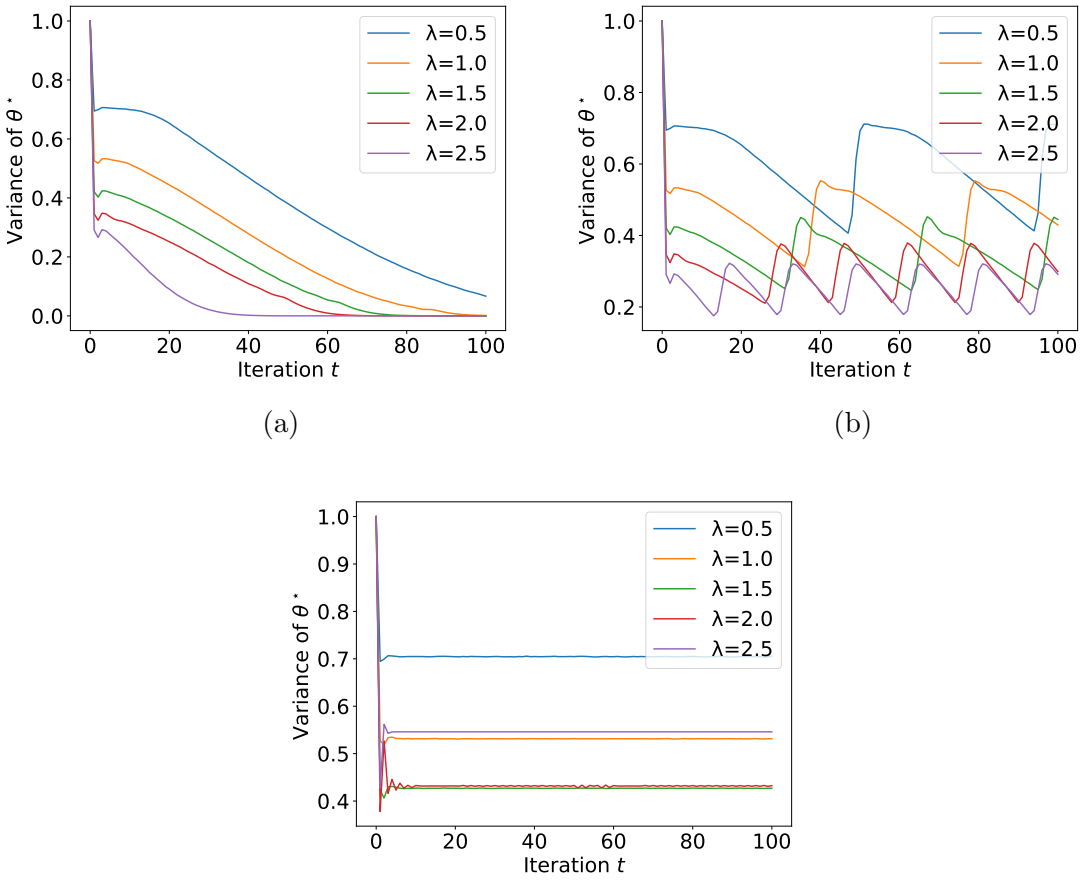


Figure 2.5: The iterative change of the variance of $\theta_A^\star$. We use $\mu_\theta = 0$, $\sigma_\theta = 1$. (a) $\Gamma = \infty$; (b) $\Gamma = 10$; (c) $\Gamma = 2$. A full simulation description is provided in Appendix A.3.

indicating that facilitating human-AI interaction can slow down the homogenization death spiral. A small λ acts as a counterforce against the death spiral, encouraging users to share more information with the AI, thereby increasing the diversity of outputs, as discussed in Theorem 2. When Γ is finite (see Figure 2.5 (b)), the death spiral phenomenon still appears initially, making it increasingly difficult for users to reduce the fidelity error. As the fidelity error accumulates, the users with very unique preferences experience significant utility loss and eventually choose to do the work themselves, partially recovering the output diversity. This recovery in diversity subsequently changes the AI prior and reduces the fidelity error for the other users, further enhancing output diversity in subsequent iterations. However, once the diversity of outputs is sufficiently restored, the users tend to rely on the AI again, causing the homogenization death spiral to reoccur. These cyclic patterns repeat over time. Moreover, when Γ is extremely low (see Figure 2.5 (c)), a lot of users opt to do the work manually, maintaining a constant level of output diversity. Nonetheless, the diversity of outputs is still less than the original because some users continue to accept the default output. Additionally, notice that in this case, increasing λ initially exacerbates the death spiral and reduces the output variance at each iteration. However, once λ becomes sufficiently large, further increases lead to more users abandoning the AI and doing the work manually, partially restoring the output diversity. This observation aligns with our findings in Section 2.4.2. Figure 2.5 confirms that the theoretical results from Section 2.5.2 remain valid in our original setting with a normal distribution of θ. The death spiral exists and the output diversity decreases over time, but promoting efficient human-AI interactions and lowering the cost of manual work can effectively mitigate the death spiral.

In what follows, we further test the robustness of our results in more complex scenarios. Specifically, we examine two additional cases. First, we explore the situation where the decision to use the AI is made retrospectively rather than prospectively. Second, we investigate scenarios where the distribution of users' preferences is more complex than a normal distribution.

Ex-post decision of accepting the AI output In the original model presented in Section 2.3, we focus on the situation where the users make an ex-ante decision about whether to use the AI to assist their work. These decisions are prospective and based on the expectation of utility loss. However, in other scenarios, users might make more myopic decisions, choosing whether to use the AI after seeing its output, which is an ex-post decision. That is, after observing the AI output θ_A and the realized fidelity error, $(\theta - \theta_A)^2$, the user will decide to accept the AI output if the realized fidelity error is less than the fixed utility cost Γ. Otherwise, the user will ignore the AI output and do the work manually. Hence, the output $\tilde{\theta}$ chosen by a user θ is:

$$\tilde{\theta} \triangleq \begin{cases} \theta_A|(\theta, \sigma_q) & \text{if } (\theta - \theta_A)^2 \leq \Gamma \\ \theta & \text{otherwise} \end{cases}.$$

In addition, since the user decides σ_q prior to deciding whether to accept the AI output, she must evaluate the expected fidelity error by considering the potential future acceptance of the AI output:

$$\tilde{e}(\theta, \sigma_q) \triangleq E[(\tilde{\theta} - \theta)^2|\theta].$$

So the utility loss of interacting with an AI becomes

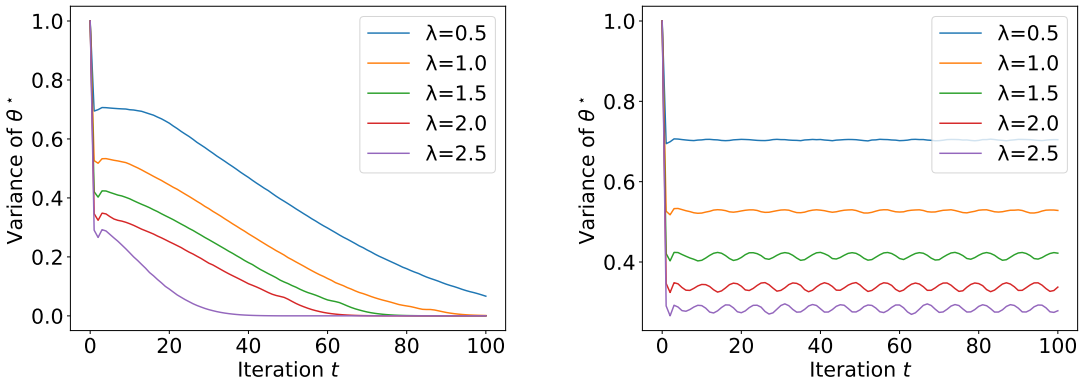
$$\tilde{l}(\theta, \sigma_q) \triangleq \tilde{e}(\theta, \sigma_q) + \lambda I(\sigma_q).$$

A user θ chooses an optimal $\tilde{\sigma}_q^\star(\theta)$ to minimizes her utility loss,

$$\tilde{\sigma}_q^\star(\theta) \triangleq \arg\min_{\sigma_q \geq 0} \tilde{l}(\theta, \sigma_q).$$

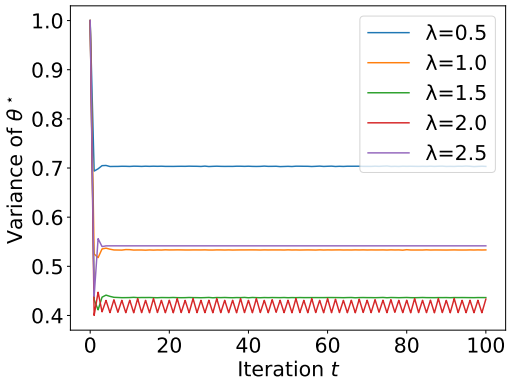
This different sequence of decisions further complicates the theoretical analysis due to the increased difficulty of solving the optimization problem. However, we demonstrate that our results remain valid through a numerical study, as illustrated in Figure 2.6.

Figure 2.6 confirms our theoretical results in this specific setting. It demonstrates that the diversity of outputs continues to decrease over time when everyone relies on the AI.



(a)

(b)



(c)

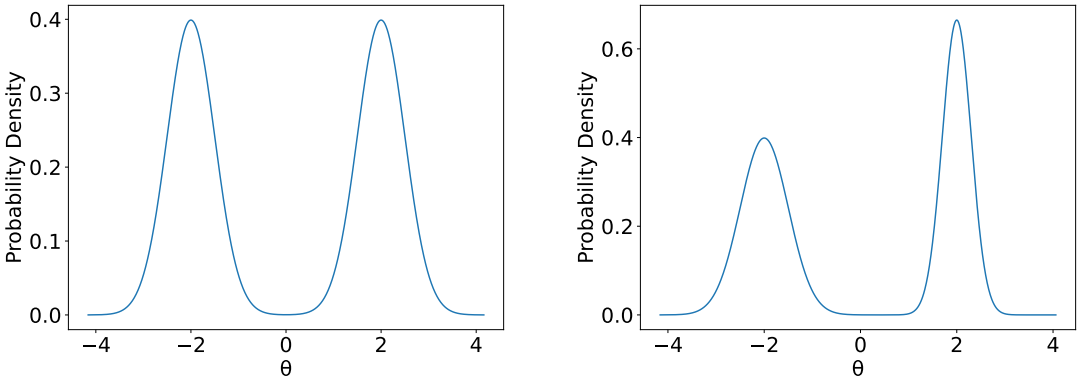
Figure 2.6: The iterative change of the variance of $\theta_A^\star$ with an ex-post decision of accepting the AI output. We use $\mu_\theta = 0$, $\sigma_\theta = 1$. (a) $\Gamma = \infty$; (b) $\Gamma = 10$; (c) $\Gamma = 2$.

However, facilitating human-AI interaction and encouraging users to perform more manual work effectively counteract the homogenization death spiral. Notably, when $\Gamma = 10$, the variance change in Figure 2.6 is less pronounced than in Figure 2.5. This is because, with ex-post decisions, the users tend to abandon the AI output earlier, rather than continuously accepting it until the expected fidelity error has significantly accumulated. As a result, the changes in the variance of outputs are less dramatic over time.

Other population distribution of users' preferences To further test the robustness of our results, we implement numerical studies using different population distributions of users' preferences. Specifically, we consider three additional types of distributions: uniform, a distribution with two symmetric peaks, and a distribution with two asymmetric peaks. The uniform distribution represents an extreme case where every preference has the same density in the population, meaning that there is no majority preference. A distribution with two symmetric peaks features two large groups of people whose preferences are on opposite sides and have the same density, potentially causing mutual reactions that alter the overall distribution of generated outputs. In contrast, a distribution with two asymmetric peaks also has two large groups of people with preferences on opposite sides, but the preferences in one of the groups are more concentrated while the other group's preferences are more diverse. This implies that one group has more homogeneous preferences, whereas the other group's preferences are more varied. The instances of the last two distribution types are illustrated in Figure 2.7.

We present the numerical results in Figure 2.8 and Figure 2.9. Regardless of the assumed distribution of θ, our insights remain consistent. The diversity of outputs continues to diminish over time when everyone uses the AI. However, a low λ or a low Γ can effectively mitigate the homogenization death spiral.

Our findings about the homogenization death spiral offer a different perspective than the technical explanation for the homogenization problem in Shumailov et al. (2023) (they



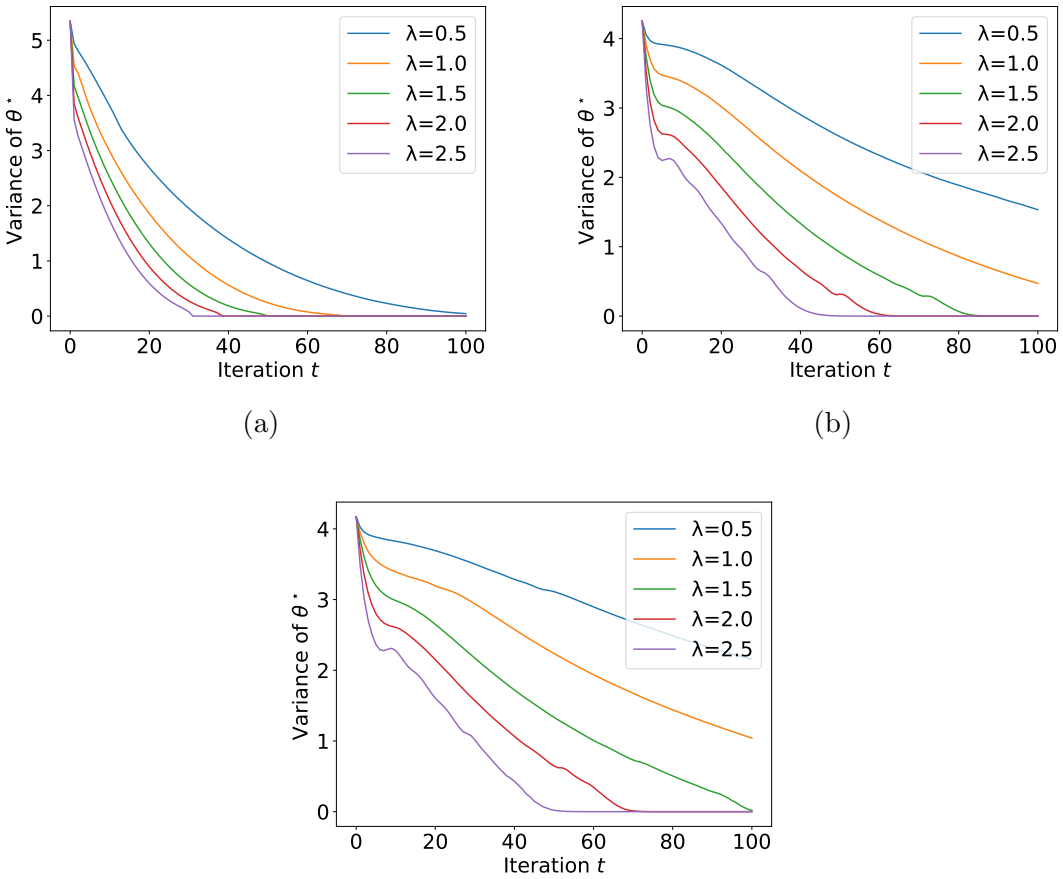
(a) (b)

Figure 2.7: The last two extra distributions in the robustness test. (a) a mixed distribution between $N(-2, 0.5)$ and $N(2, 0.5)$; (b) a mixed distribution between $N(-2, 0.5)$ and $N(2, 0.3)$.

call it model collapse), where the authors suggest the problem is caused by sampling and approximation errors. Our model also indicates that human and technical factors may mutually reinforce each other, potentially leading to an exacerbated homogenization problem. In our model, the homogenization loop is due to technical factors, the misspecified AI prior, and human behavior, who maximize their utility and are willing to let go of their specificity to limit the communication costs. We also offer a solution: creating models facilitating human-AI interactions (i.e., low λ) can significantly slow down the homogenization process.

2.6 Human-AI Interactions and AI Bias

The homogenization phenomenon shows that the use of AI "influences" the user outputs, in the sense that $\theta^\star \neq \theta$ for many users θ. This is potentially concerning, as any choices made in the AI training, any bias it might have, would then influence the users' choice of output. Indeed, generative AIs are not necessarily trained to reflect the population's preferences

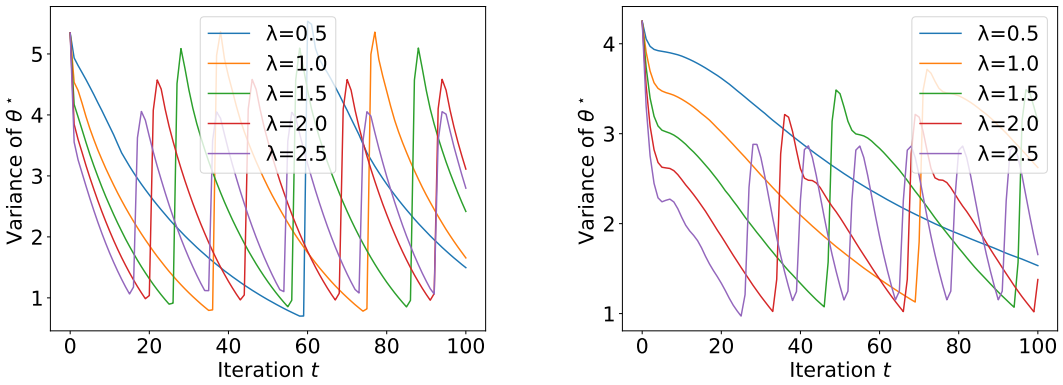


(a)

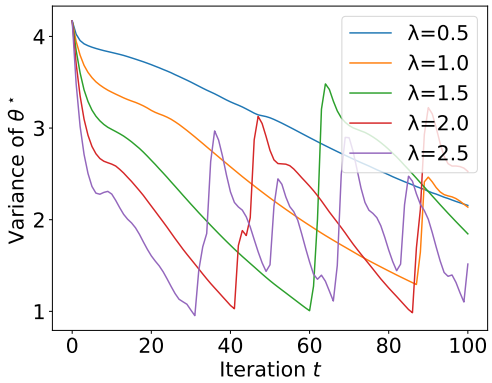
(b)

(c)

Figure 2.8: The iterative convergence of the variance of $\theta_A^\star$ in the three cases with a more complex distribution of θ when $\Gamma = \infty$. (a) uniform; (b) a mixed distribution between $N(-2, 0.5)$ and $N(2, 0.5)$; (c) a mixed distribution between $N(-2, 0.5)$ and $N(2, 0.3)$.



(a) (b)



(c)

Figure 2.9: The iterative change of the variance of $\theta_A^\star$ in the three cases with a more complex distribution of θ when $\Gamma = 10$. (a) uniform; (b) a mixed distribution between $N(-2, 0.5)$ and $N(2, 0.5)$; (c) a mixed distribution between $N(-2, 0.5)$ and $N(2, 0.3)$.

exactly. For example, the AI's training data may be censored to avoid illegal or dangerous behavior (Thompson, 2023). Moreover, the training of LLMs uses Reinforcement Learning from Human Feedback (Ziegler et al., 2020), in which a small group of humans "teach" the model what output is preferable. These training choices of a few can then influence the output of the entire population interacting with AI. We model this potential AI "bias" via an AI prior that does not exactly reflect the population's preference distribution (i.e., $\mu_A \neq \mu_\theta$ or $\sigma_A \neq \sigma_\theta$), leaving the true user preference distribution and the rest of the Bayesian inference unchanged. We refer to $\mu_A \neq \mu_\theta$ as a *directional bias* and to $\sigma_A < \sigma_\theta$ as a *censoring bias*. In Example 1, the AI may have a slight bias towards a political side (directional bias), or it may avoid extreme political views (censoring bias). We first discuss how the two types of AI bias affect users and the effectiveness of human-AI interactions. We then evaluate how much influence a biased AI can have on society and ways to mitigate this influence.

2.6.1 AI Bias and User Utility

A biased AI may be less useful for some users but may also help others, as summarized below.

Proposition 4. The utility loss $l^\star$ of a user θ changes when with a biased AI as follows:

1. the directional bias favors users the AI is biased towards: $l^\star$ strictly increases with $|\mu_A - \theta|$;
2. the censoring bias benefits users with common preferences: $l^\star$ strictly increases in σ_A when $\sigma_A \geq |\mu_A - \theta|$, and strictly decreases in σ_A when $\sigma_A < |\mu_A - \theta|$.

Item 1 of the proposition states that directional bias is detrimental to users of the "opposite" direction. In Example 1, if the AI is slightly right-leaning, a left-leaning journalist may need more communication cost to obtain an article they will be satisfied with. However, a

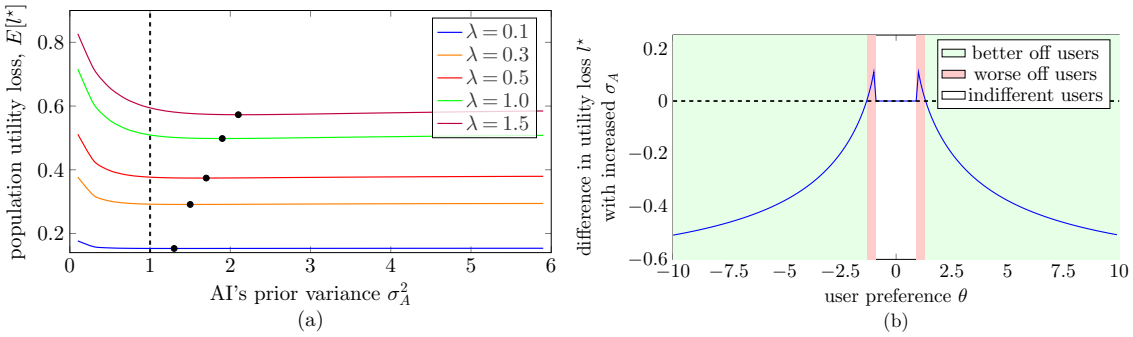
right-leaning journalist could be directly satisfied with the default output. The ideal case for user θ is $\mu_A = \theta$, as the default AI output would correspond to a perfect utility $l^\star = 0$. Item 2 states a similar result for the censoring bias. To clarify it, suppose $\mu_A = \mu_\theta$, and consider a user with "common" preferences less than a standard deviation away from the mean, i.e., $|\mu_\theta - \theta| < \sigma_\theta$. Then she would be better off if a slight censoring is used, with σ_A such that $|\mu_\theta - \theta| < \sigma_A < \sigma_\theta$. When reducing σ_A, the AI is more likely to return outputs closer to the mean, benefiting this user. However, this hurts users with more unique preferences, who will need more communication costs to maintain a reasonable fidelity or will stop using the AI. Therefore, both types of bias can increase some users' utility loss and decrease others'. The next results consider the aggregate-level consequences of bias and its effect on the population utility, defined as the expected utility loss $E[l^\star]$ taken across the users θ.

Proposition 5. Directional and censoring bias have contrasting effects on the population utility:

1. A small directional bias has a limited negative effect on the population utility: $\left.\frac{\partial E[l^\star]}{\partial \mu_A}\right|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} = 0$ and $E[l^\star]$ is minimized at $\mu_A = \mu_\theta$.
2. A small censoring bias can have a stronger negative impact: for example, $\left.\frac{\partial E[l^\star]}{\partial \sigma_A}\right|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} < 0$ when $\lambda \geq 2\sigma_\theta^2$ and $\Gamma \to \infty$.

The proposition first shows that, while any directional bias hurts the population utility, a small directional bias has a negligible effect. Intuitively, if $\mu_A = \mu_\theta + \varepsilon$ for $\varepsilon > 0$ small, slightly less than half of the users (above $\mu_A + \epsilon/2$) benefit from the bias because they have a lower communication cost for the same fidelity, while the other half (below μ_θ) is hurt because of an increased communication cost for the same fidelity. These two populations balance each other, which limits the total loss of utility.

The case of censoring bias (Item 2 of Proposition 5) is maybe more surprising. Unlike the effect directional bias, setting $\sigma_A = \sigma_\theta$ (an unbiased prior) does not generally minimize



(a) (b)

Figure 2.10: (a) the black circles indicate the AI prior variance σ_A^2 that would minimize the population utility loss. (b) the utility loss $l^\star$ with $\sigma_A^2 = 2$ minus those with $\sigma_A^2 = 1$, when $\lambda = 1$. In both panels, we use $\mu_A = \mu_\theta = 0, \sigma_\theta^2 = 1, \Gamma = +\infty$.

the population utility loss $E[l^\star]$. Both the proposition and Figure 2.10 (a) show that when $\Gamma \to +\infty$, it is preferable to have $\sigma_A > \sigma_\theta$ (the opposite of censoring). Remember from Section 2.4.1 that there are two types of user behavior when $\Gamma \to +\infty$: using the default AI output or interacting with the AI, and the choice of σ_A only influences the utility of the interacting users. Therefore, the AI Bayesian update is more accurate when the choice of σ_A better represents the interacting users, who are the ones with more unique preferences (Proposition 1). This is why choosing $\sigma_A > \sigma_\theta$ improves the population utility. This effect is illustrated in Figure 2.10 (b): when increasing σ_A, common-preference users do not lose utility, but more unique users see a large improvement in utility loss. While this result may have implications for the design of interactive AI, it also warns against the potential negative effects of censoring bias. Decreasing σ_A is particularly hurtful to the most unique users, *who rely on human-AI interactions the most.* While censoring can be useful in preventing dangerous or illegal uses of AI, our results also highlight the importance of training AI on datasets that reflect a wide range of preferences.

2.6.2 AI Bias Becomes Societal Bias

Another interpretation of Item 1 of Proposition 5 is that a small directional bias $|\mu_A - \mu_\theta| > 0$ (referred to as *AI bias* in this section) may be hard to detect in practice, as it does not strongly affect the population's utility. However, it may still significantly influence the user output $\theta^\star$. For example, users who accept the default output ($I^\star = 0$) have $\theta^\star = \mu_A$, directly inheriting the AI bias. On the other hand, users may choose to share more information to correct this bias and maintain a high-fidelity output. To study which effect dominates, we analyze the consequences of the AI bias on the *societal bias*, defined as the bias of the output distribution: $|E[\theta^\star] - \mu_\theta|$.

Theorem 3. *Given the AI bias $|\mu_A - \mu_\theta|$ and the societal bias $|E[\theta^\star] - \mu_\theta|$,*

1. *the societal bias is lower than the AI bias,*
2. *the societal bias is minimized when $\lambda \to 0$ or $\Gamma \to 0$: $|E[\theta^\star] - \mu_\theta| = 0$,*
3. *the societal bias is maximized when $\lambda \to +\infty$ and $\Gamma \to +\infty$: $|E[\theta^\star] - \mu_\theta| = |\mu_A - \mu_\theta|$,*
4. *if everyone uses AI ($\Gamma \to +\infty$), the societal bias increases with the cost of human-AI interactions λ.*

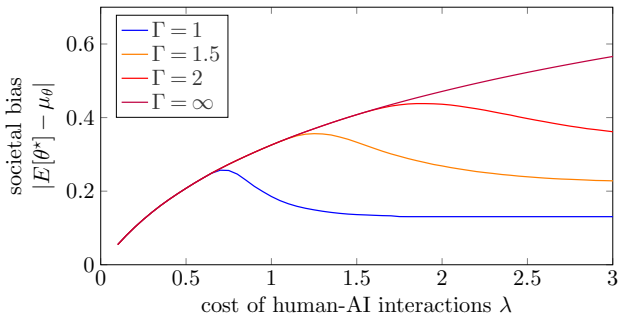


Figure 2.11: We use $\mu_\theta = 0$, $\mu_A = 1$, $\sigma_\theta = \sigma_A = 1$ (e.g., the AI bias is $|\mu_A - \mu_\theta| = 1$).

This theorem is illustrated in Figure 2.11 and shows an encouraging result: human-AI interactions can partially prevent AI bias from becoming societal bias. For example, a left-

wing journalist in Example 1 may increase their interactions with the AI to correct the output if the AI is biased to the right. This is particularly true when either the cost of human-AI interactions, λ, or the cost of not using AI, Γ, is low. It is much easier for users to correct bias if they can easily interact or simply stop using AI. However, Theorem 3 also comes with a warning. For larger tasks that are painful to do without AI (high Γ), and if the human-AI interactions are not efficient (high λ), rational users will simply accept the AI bias, which will be fully converted into a societal bias. For example, generative AI systems that favor speed over interactivity (e.g., the AI writing assistant Grammarly) or tackle complex tasks with limited interactivity (e.g., the image generator Midjourney) may fall into this category. Any bias they introduce may have a stronger influence on societal output than systems designed for interactivity (e.g., ChatGPT).

2.7 Conclusions

The widespread introduction of generative AI enables significant productivity gains. However, we show that the power of these tools may lead users to accept homogenized or biased outputs and abandon their idiosyncrasies, even when given the possibility to communicate their preferences. At the societal level, this can lead to homogenization (reinforced by training loop effects) and the potential influence of AI training choices on the societal output. These risks are particularly strong for labor-intensive tasks (e.g., image/sound generation) or with AI tools that favor speed over preference-sharing (e.g., grammar correction). Nonetheless, we also show that enabling easier human-AI communication and training the AI on diverse data can significantly limit these negative effects, allowing the best of both worlds: high productivity and human diversity.

The topic studied in this work combines technical and behavioral complexity, as we need to capture how the AI tool works and how users interact with it. Our Bayesian framework is a simplified representation of this interaction that still enables nontrivial insights, but

there are effects we do not capture. For example, it is a simplification to assume that a one-dimensional normal distribution can represent the vast space of human preferences and outputs and that the complexity of human-AI communication can be represented as a simple normal signal and Bayesian inference. We also assume all users have the same no-AI utility loss Γ, and the same human-AI interaction cost λ for a given task. Nonetheless, we believe our framework is versatile enough to study deeper variants and is a first step towards understanding the societal consequences of human-AI interactions.

Recent empirical studies examine the multifaceted implications of generative AIs across various domains, such as education (Baidoo-Anu and Owusu Ansah, 2023), labor markets (Eloundou et al., 2023), and marketing (Brand et al., 2023). Understanding the general effects of user behaviors while interfacing with an AI remains an open question that is difficult to study empirically. We hope our analytical approach highlights the importance of adopting a human-centric perspective rather than solely focusing on AI technology. Indeed, while AIs could surpass human abilities in various aspects (Binz and Schulz, 2023; Chen et al., 2023; Webb et al., 2023), their impact may largely depend on how we employ them. The interaction with AIs could offer a novel medium for production and creation but also introduce an extra risk: AIs may filter and even replace our original preferences, styles, and tastes, thereby leading to a world dictated by the AI creators' perspective — a potentially homogenized and biased world. Improving human-AI interactions and encouraging users to authentically voice their unique views is crucial to avoid these societal pitfalls.

CHAPTER 3

Autonomous Vehicles in Ride-Hailing and the Threat of Spatial Inequalities

3.1 Introduction

Ride-hailing platforms such as Lyft, DiDi, and Uber have become an integral part of urban transportation systems. In New York City, for example, these platforms average an impressive 700K daily trips and 80K unique drivers per month.[1] This is not a surprise: commuters can now go seamlessly from point A to point B using their smartphones. Drivers have control over their schedules, and, more importantly, anyone can be a driver. This has brought an unprecedented but well-grounded sense of reliability among users, who almost certainly will get a ride upon request no matter where they are.

Meanwhile, advances in the development of autonomous vehicles (AVs) are gathering interest, foreshadowing a fundamental change in the ride-hailing industry. AVs may lead to a substantial decrease in operating cost per mile (Fagnant and Kockelman, 2018; Hazan et al., 2016; Litman, 2023) and can be controlled centrally, which improves reliability. This has led several platforms, such as DiDi and Lyft, but also others, such as Google and Amazon, to invest in self-driving car technology (DiDi, 2023; Lyft, 2024; Uber, 2024). Uber signed a 10-year deal with Motional (Davalos, 2022) to pair Motional's AV technology with Uber's delivery and ride-hailing platform. Amazon's Zoox started testing its ride-hailing service

[1]Source: NYC TLC data. `https://www1.nyc.gov/assets/tlc/downloads/pdf/2020-tlc-factbook.pdf`

on the streets of Las Vegas in June 2023 (Ludlow, 2023). Furthermore, Waymo is already operating a ride-hailing service with fully autonomous rides in Phoenix, AZ, San Francisco, CA, and Los Angeles, CA (Waymo, 2024).

As platforms introduce AVs more broadly in cities, they will likely face limitations that can prevent them from switching to a fully autonomous service. For example, AVs can only operate in areas with highly detailed mapping; they can also be subject to weather limitations. The significant investment this new technology requires may also limit the number of vehicles available (Litman, 2023). Furthermore, some riders may also simply prefer HVs to AVs. Therefore, as we transition to a fully autonomous world, the ride-hailing industry will likely change to an operational mode characterized by an augmented fleet composed of a mix of HVs and AVs. Indeed, Lyft anticipates that their first generation of self-driving service will be deployed on a hybrid AV network that includes human drivers (Hur, 2022). Uber and Waymo have created a partnership that makes it possible for Uber riders to hail a fully autonomous Waymo ride within the Uber App in Phoenix, AZ.[2]

The different economic structures of these two types of supply, coupled with the trade-off between a platform's profitability and service reliability (e.g., having a low wait time and a high match rate), raise the question of how operating a mixed fleet can impact riders and drivers. First, the participation of human drivers must be secured through sufficiently high incentives and pay. Second, while autonomous vehicles do not need to be incentivized, the platform must incur the cost of buying and maintaining them. In principle, as a platform introduces AVs, it may prefer to lean more on them because they are a sunk cost, and using them is potentially more profitable than paying for HVs. However, AV deployment can impact human participation, which, in turn, risks negatively affecting reliability. This potential impact on reliability may affect different regions in a city differently, creating spatial inequality of access to transportation. If the AVs primarily focus on high-demand areas like

[2]link accessed on May 18th 2024: https://waymo.com/blog/2023/10/the-waymo-driver-now-available-on-uber-in-phoenix/

downtowns and airports, the reduction in human participation could particularly negatively affect low-demand regions such as the outskirts of a city.

In this study, we aim to shed light on the impact of the introduction of self-driving cars in the fleet of a ride-hailing platform. We are interested in how the coexistence of AVs and HVs, as well as the gradual adoption of AVs by the platform, may affect the platform's supply management and key operational metrics. In particular, we seek to elucidate the impact of HVs participation decisions and the subsequent consequences on service levels and spatial equality in access to transportation across different regions.

3.1.1 Main Contributions

We first develop a game-theoretical queueing model that parsimoniously captures key aspects of the mixed-fleet management problem faced by a platform. This model enables us to illustrate the potential consequences of introducing AVs. While we limit the complexity of the model to enable formal analysis, we then validate and extend our findings using a highly detailed and more realistic simulation using New York City data. In the model, the platform acts as a leader that determines the distribution of its fleet across locations and, at each location, decides how to match new requests with either AVs or HVs to maximize its profit rate. Human drivers are the followers who, given the platform's policy, decide whether to join the platform by gauging their potential earning rates against an outside option. A wage equilibrium, therefore, determines the number of HVs participating. We focus on the effects of operational controls by assuming that market prices are fixed, the per-trip and vehicle-type profits are exogenously given, and AVs are more profitable than HVs. However, we relax these assumptions later in the simulation study.

We uncover a paradoxical effect: *the introduction of autonomous vehicles may deteriorate the platforms' reliability.* Because the marginal cost of operating HVs is higher (drivers need to be paid), a profit-maximizing platform should prioritize AVs, i.e., prefer to assign new requests to AVs rather than HVs. However, this decreases the earnings of HVs, which

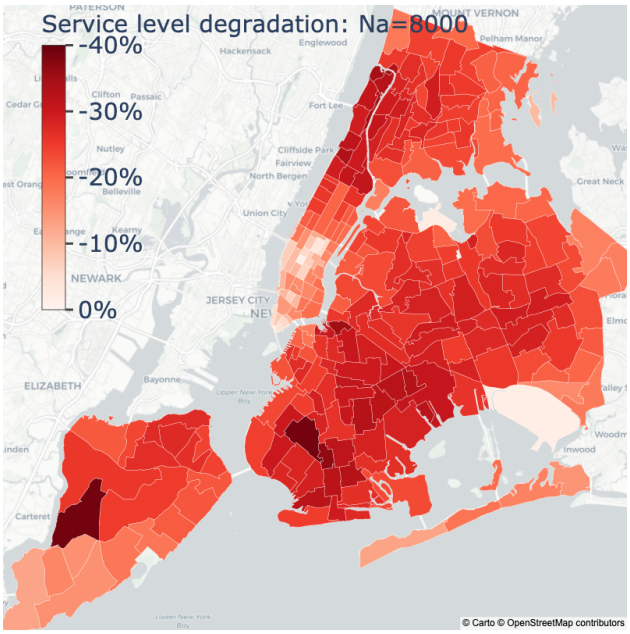


Figure 3.1: Estimates of service level degradation if 8,000 AVs were introduced in New York City, based on a detailed vehicle-level simulation. Darker shaded areas experience a higher service degradation. Service level degradation is measured as the relative decrease in the number of successfully served trips. Note that suburban areas are much more affected than central areas.

incentivizes more human drivers to leave the platform than otherwise would have without prioritization which causes an overall service level decline. As the platform introduces more AVs, the service level continues to decrease until it is no longer profitable for the platform to prioritize its AV fleet. The newly added AVs are then treated equally to human drivers, and the service level remains unchanged until the supply of AVs drives all the HVs out of the market. Then, the service level begins to recover as the platform increases its AV fleet. We also uncover some cases where the platform should not prioritize AVs and should prioritize HVs instead. This happens when the driver-pay ratio is small (i.e., the platform's commission rate is high), although our simulations show this is not likely to happen in practice.

We also demonstrate that the service level degradation described above is *not homoge-*

neous across locations in a city. In particular, high-demand locations are less affected than low-demand locations, and this discrepancy increases when the platform introduces (and prioritizes) more AVs. The reason behind these findings is that an AV prioritizing policy not only prioritizes AVs in dispatch but also tries to maintain them in areas with higher demand, pushing de-prioritized vehicles to lower-demand areas. Indeed, it is more profitable to operate AVs with high utilization, which is achievable in high-demand areas. For example, if a request goes to the outskirts of a city, it would not be efficient to assign it to an AV as it would likely have to come back empty: the platform would, therefore, choose dispatch and relocation policies that try to maintain the AVs in high-demand areas. Consequently, the earnings of HVs are reduced not only because of the additional number of AVs in the market but also because they are deprioritized and relegated to less profitable areas.

To show that the learnings from this theoretically tractable model are robust to more realistic settings, we designed a simulation study in New York City. Using publicly available historical ride-hailing data in NYC, we simulate the fleet management problem of a ride-hailing company introducing AVs in NYC. The simulation retains the critical driving features of the queueing model, such as a profit-maximizing platform and HVs joining decisions. However, to approximate a real-world scenario, it also adds many features that are absent in the model. For example, vehicles are simulated individually inside the city network, and we use a state-of-the-art relocation policy when they are idle. Customers' preferences are also modeled in detail with a random utility model, and can choose whether to request an AV or an HVs. However, the platform can still influence the customer's choice between AVs and HVs by using a differentiated pricing policy. Despite the modeling differences, we find results that are consistent with our main results. Urban areas with higher demand benefit the most from the introduction of AVs, experiencing improved service levels. In contrast, suburban areas with lower demand experience a degradation in service levels (see Figure 3.1).

Furthermore, the simulation delivers additional insights that are only available in a more granular spatial setting. First, as the platform introduces AVs, the average ETA (passenger

wait time) increases significantly and disproportionately in more remote areas, which is consistent with service level changes. Also, we observe that the service level may be affected by other spatial factors, such as distance between areas and demand imbalance. The service level of a low-demand area can be positively influenced if this area is closer to high-demand areas or has more incoming than outgoing trips. As a result, the most seriously affected areas are those that have low demand, receive fewer incoming trips, and are far from high-demand regions.

Our work shows that pure profit maximization when introducing AVs could lead to undesirable and maybe unexpected social outcomes, suggesting that careful regulation should be considered.

3.1.2 Related literature

Related studies on autonomous vehicles. Numerous studies consider various challenges in technology, safety, travel behavior, public transportation, environment, and governance, among others, that could arise in society due to the presence of autonomous vehicles. We refer the reader to the surveys by Hussain and Zeadally (2019), Ma et al. (2020), and Narayanan et al. (2020) for excellent discussions of these aspects. The common denominator in these works is their focus on autonomous vehicles rather than a mixed-fleet system with both autonomous vehicles and strategic human drivers. Some of these works describe the potential benefits that AVs can bring to society. For example, Fernandes and Nunes (2010) suggest that the platooning of AVs could cause a significant increase in capacity for both urban roads and highways. Similarly, Mirzaeian et al. (2021) use a queueing model to demonstrate the potential congestion reduction on multi-lane highways. Additionally, Baron et al. (2022) investigate the impact of AVs on social welfare in terms of connectivity, comfort, and collaborative consumption in a system in which some households free their driving time by using AVs while others spend time driving. And Reed et al. (2022) demonstrate that autonomous vehicles can improve package delivery services. With the help of an autonomous-assisted

system, a delivery person may spend less time searching for parking or walking back to a parking spot.

In this study, we depart from the studies mentioned above by considering a mixed-fleet ride-hailing platform that owns an AV fleet and has to set proper incentives for HVs to join the market. We show that introducing AVs may hurt accessibility to transportation as measured by a city-wide service level reduction that affects suburban areas more negatively. A primary distinguishing factor of our work is a wage equilibrium constraint that ensures driver earnings are high enough.

Only a few attempts have been made to analyze the impact of AVs in the ride-hailing market. Siddiq and Taylor (2022) study the influence of AVs on the competition between two ride-hailing companies, where only one of the companies has access to AVs. The authors analyze how access to AVs affects platform profit, agent welfare, and social welfare. Lian and van Ryzin (2022) consider a market with different dispatch platform designs (common vs. independent) and different AV competition levels (monopoly vs. competitive). They investigate how various market designs affect the equilibrium price and social welfare. The authors establish that a common market with both AVs and HVs is plausible because AVs with high fixed costs are less flexible and may not serve all demands cost-effectively. They also show that the lower operating cost of AVs may not necessarily lead to a lower-price ride-hailing market. Ostrovsky and Schwarz (2019) discuss the interdependence between AVs, road pricing, and carpooling. They demonstrate that due to the overall progress in information technology, AVs can make road pricing and carpooling more convenient and attractive. Lanzetti et al. (2023) analyze the equilibrium outcome in terms of market share in a system with multiple competing transportation modes, including AVs. They show that depending on policy constraints and market conditions, a ride-hailing system with AVs may benefit or harm public transportation in terms of market share. Mirzaeian et al. (2021) studies the impact of AVs on highway congestion through a queueing model and compares two policies: a designated-lane policy for AVs and an integrated policy with mixed traffic.

In contrast to these studies, our work focuses on the change of operational decisions (e.g., matching) of the platform when introducing AVs and their consequences on service levels at a system-wide level and a more granular, location-wise level.

In the context of operational decisions, Benjaafar et al. (2023) use a fluid model to characterize the optimal re-routing decision of the platform with a mixed fleet of AVs and HVs and show that the introduction of AVs may not necessarily harm drivers. In their model, the prioritization policies always match one of two types of vehicles without considering the destination of trips or allowing partial prioritization. In contrast, our model optimizes both the matching and the allocation of vehicles by using independent queues, and our prioritization policies may be partial, meaning that we allow for a continuum of prioritization levels. And we also show that HVs may not be harmed if the pay ratio is low enough. Finally, Freund et al. (2022) adopt a sequential game model and show that AVs may be under-utilized to ensure HVs remain engaged in the market when platforms outsource AVs. To avoid an unbounded profit loss, they propose a prioritization contract that increases the utilization of AVs. Our study assumes that the platform owns AVs and also shows that if the number of AVs is large enough, it may be optimal for the platform to partially prioritize AVs to maximize profit rather than fully prioritize them all the time. Nonetheless, we optimize matching and allocation strategies and focus on the impact of such a prioritization policy on service levels in terms of system-wide and spatial variation.

Related studies on ride-hailing systems. Several papers in the ride-hailing literature have studied the problem of matching riders to drivers. Closest to our work are the papers that use a queueing modeling approach to capture the operational controls of the platform to determine matching between riders and drivers (see, e.g., Banerjee et al. (2018), Kanoria and Qian (2020), Özkan and Ward (2020), Hu and Zhou (2022)). Typically, this space focuses on determining an optimal matching policy with human drivers only. In our work, however, we assume there are multiple sources of supply with different economic structures in the system we study. In this setting, matching requires careful balancing of the earnings

of human drivers and the profit of the platform, which, all else being equal, may improve when one type of supply is prioritized over the other.

In addition, our work is broadly related to the study of how the performance of a ride-hailing market is affected by other platform controls, such as repositioning and admission control (see, e.g., Braverman et al. (2019), Afèche et al. (2023) and Wang et al. (2022)). In our queueing model, we assume the platform can directly determine the allocation of vehicles to different locations, which can be seen as an outcome implied by these levers. This allows us to focus on how the prioritization and allocation of vehicles may affect service levels. In particular, Afèche et al. (2023) show that it may be optimal for the platform to strategically reject demand at a low-demand location and induce drivers to reposition to a high-demand area. We show that the platform prefers to serve ride requests with AVs in high-demand locations. At the same time, HVs are relegated to low-demand areas, leading to the spatial inequality of service level degradation. Moreover, while our primary focus in the queueing model is on the matching and allocation policy, we also incorporate a pricing policy within our simulation. Consequently, our work is also related to studies on pricing in ride-hailing platforms (e.g., Castillo et al. (2017), Bimpikis et al. (2019), Besbes et al. (2021), Hu et al. (2022), Cachon et al. (2022)). For example, Cachon et al. (2022) examine the key trade-offs between a centralized pricing strategy managed by the ride-hailing platform and a decentralized pricing strategy determined by the drivers. Bimpikis et al. (2019) focus on spatial price discrimination on a ride-sharing platform and show that location-based pricing can significantly benefit the platform when demand is unbalanced.

Several other studies also consider the inclusion of multiple types of vehicles on ride-hailing platforms. For instance, Lu et al. (2024) explores the optimal regulatory strategy for maximizing total welfare by considering the coexistence of traditional taxi services, which operate through both street-hailing and platform-based modes, and private car drivers, who serve consumers exclusively via the platform. Additionally, Fatehi (2024) investigates the optimal policy of a profit-maximizing platform that manages a mixed fleet of electric vehicles

and traditional internal combustion engine vehicles.

Related studies on blended workforces. AVs and HVs can be seen as two labor forces with different incentive structures: AVs are owned by the platform, while HVs must be incentivized to participate in the market. In this sense, our work is also related to the emerging literature on blended workforces in the gig economy (see, e.g., Dong and Ibrahim (2020), Lobel et al. (2024), He and Goh (2022), Chakravarty (2021), Castro et al. (2022) and Hu et al. (2023)). For instance, Hu et al. (2023) addresses the classification of workers in the on-demand economy with two types of workers: full-timers and part-timers. Their analysis reveals that classifying all gig workers as employees can reduce the welfare of full-timers due to undercut by profit-maximizing companies Notably, our work shares some similarities to Krishnan et al. (2022), which studies an actual implementation of a blended workforce using prioritization. The authors explain how a ride-hailing platform should classify the drivers into "priority drivers," prioritized by the matching system to have higher earnings and regular "flexible drivers." This study also considers a blended workforce and optimal prioritization strategies. Still, our focus is on the service level impact of introducing AVs in a ride-hailing market where the platform has to determine how to treat different types of vehicles and their corresponding spatial distribution.

Methodology. In terms of methodology, we solve our multi-location queueing model using *the achievable region approach*. In the context of stochastic optimization, this approach seeks the solution to a stochastic optimization problem by identifying a feasible performance space and solving the problem within this space. For example, Bertsimas (1995) use this method to describe the achievable region of a multi-class queueing control problem as a convex polyhedron and yield a mathematical program for which efficient algorithms are available. We refer the reader to Bertsimas and Niño-Mora (1996), and Dacre et al. (1999) for additional results related to this approach. In our work, we adopt the achievable region approach to determine the feasible supply arrival rates at each location that can emerge from different dispatch policies. Once we obtain the optimal arrival rates, we can reverse-engineer

the corresponding dispatch policy (c.f., Section 3.3).

The remainder of this chapter is organized as follows. In Section 3.2, we present a game-theoretic queueing model for a hybrid system with AVs and HVs. In Section 3.3, we interpret how to solve the problem using the achievable region approach. In Section 3.4, we derive macro-level insights from the queueing model. And Section 3.5 reveals the spatial disparity of service level. In Section 3.6, we present a simulation study of New York City to illustrate that our main results still hold in a more realistic setting and further discuss some additional spatial effects observed in the simulation. We conclude in Section 3.7.

3.2 Model Setup

In this section, we formulate the problem faced by a ride-hailing platform that operates a mixed fleet of human and autonomous vehicles using a game-theoretical queueing model. This model is simple enough to obtain precise results yet rich enough to capture the fundamental trade-offs that emerge in this mixed fleet setting. We note that our goal here is to introduce a parsimonious model that enables us to reveal insights rather than capture all the nuances of a ride-hailing system. Later in Section 3.6, we demonstrate the robustness of our findings in the queueing model in a large-scale simulation using NYC data.

Trip requests. Requests for trips arrive to a city with locations $j \in \{1, \ldots, L\}$ according to a Poisson process with rate $\boldsymbol{\mu} = \{\mu_j\}_{j=1}^{L}$, where L is the number of locations in the city. The average duration of trips departing from location j is τ_j. We consider impatient riders who will leave the system immediately if, upon arrival, they are not assigned a ride. For example, a rider may be able to find a ride on a different platform.

AV and HV fleets. We let N_A and N_H be the steady-state average numbers of AVs and HVs that operate in the city, including both idle vehicles and those serving requests. N_A is considered as an exogenous quantity, and our goal is to study how changes in N_A will affect the platform's service levels across locations. (In Section 3.4 we discuss the choice

of the optimal N_A when it is a tactical decision the firm must make.) On the other hand, N_H is set endogenously as drivers decide whether to join or exit the platform, following a wage equilibrium that we formally introduce below (c.f., Equation (3.4)). We model vehicles becoming available at each location j as independent Poisson processes with arrival rates $\boldsymbol{\lambda}_A \triangleq \{\lambda_{A,j}\}_{j=1}^L$ for AVs and $\boldsymbol{\lambda}_H \triangleq \{\lambda_{H,j}\}_{j=1}^L$ for HVs. They queue at the location and are dispatched to incoming requests following dispatch policies set by the firm. We next discuss how these arrival rates are set. Figure 3.2 illustrates this queueing setup.

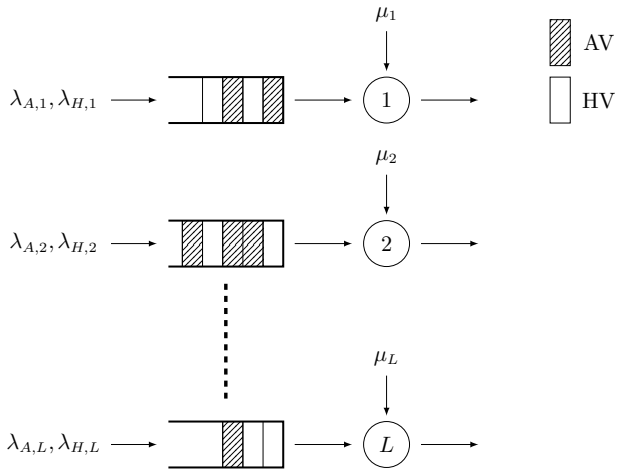


Figure 3.2: Queueing Model Structure

Dispatch policy and vehicle balancing. Let Π denote a set of dispatch policies $\pi \in \Pi$ the platform can use. Π is general, and we allow any dispatch policy: they may be idling, they may vary by location, they may prioritize one vehicle type over the other, and they may be randomized. A dispatch policy π then induces an expected steady-state waiting time function $W_{i,j}^{\pi}(\lambda_{A,j}, \lambda_{H,j})$ for vehicle type i at location j (Section 3.3 derives an explicit expression). We will assume that the platform can perfectly control vehicle distribution over locations. While this control is more limited in practice, the combination of pricing, dispatch, and relocation policies are levers platforms can use to influence the vehicle's spatial distribution significantly — as shown in our realistic simulations. Formally, we assume the platform can balance the vehicles across locations without limitation as long as the total

number of vehicles is maintained. That is, for given N_A and N_H, the platform can choose vehicle arrival rates $\lambda_{A,j}$ and $\lambda_{H,j}$ for each location j that are consistent with their respective average number of vehicles (via Little's Law). Let $N_{i,j}$ denote the steady-state average number of vehicles of type i in location j. $N_{i,j}$ depends on the vehicle arrival rates and the dispatch policy and can be obtained from Little's Law as:

$$(\tau_j + W^{\pi}_{i,j}(\lambda_{A,j}, \lambda_{H,j})) \cdot \lambda_{i,j} = N_{i,j}, \quad i \in \{A, H\},\ j \in \{1, \ldots, L\}, \tag{3.1}$$

where $\tau_j + W^{\pi}_{i,j}(\lambda_{A,j}, \lambda_{H,j})$ is the expected time that vehicles of type i in location j spend waiting and then serving a customer. The platform can choose any $\boldsymbol{\lambda}_A$, $\boldsymbol{\lambda}_H$, and π as long as the total number of vehicles of each type across locations is maintained:

$$\sum_{j=1}^{L} N_{i,j} = N_i, \quad i \in \{A, H\}, \tag{3.2}$$

Note that Equation (3.2) is the only constraint that affects the locations jointly, as it represents the need for the platform to balance its fleet across the locations. Note that our model can be interpreted as a *fixed-population-mean model*, which approximates the dynamics of a closed network of queues. In a closed network, given the number of vehicles, the routing probabilities, and a matching and relocation policy, the arrival rates of vehicles to each location are determined. A fixed-population-mean model approximates such a network by an open network of queues with Poisson arrivals in which the arrival rates are such that they make the average number of vehicles (overall and at each location) consistent with the exact number of vehicles in the closed queueing network. We refer the reader to Whitt (1984) for more detail about this approach and approximation guarantees. In Section 3.6, we build a simulation that explicitly captures vehicles' dynamics and does not assume perfect vehicle balancing.

Service level. The service level at location j, ρ_j, is given by

$$\rho_j = \frac{\lambda_{A,j} + \lambda_{H,j}}{\mu_j}, \quad j \in \{1, \ldots, L\}. \tag{3.3}$$

This corresponds to the fraction of requests that the platform can serve at location j.

Revenue and profit. We assume that the revenue from each human driver trip is shared between the platform and the driver. Human drivers keep a fraction $\gamma \in (0,1)$ of the revenue of the trip,[3] whereas the platform keeps all the revenue from each AV trip. Nonetheless, the platform has to pay a capital cost of AVs, C_A, per hour, and an additional hourly operational cost of AVs, c_A, for each AV trip. Let $P_A > 0$ and $P_H > 0$ denote the hourly price rate from each AV and HV trip, respectively.[4] Then, the platform's hourly profit at location j stemming from AVs is given by $(P_A - c_A) \cdot \tau_j \cdot \lambda_{A,j}$, while its hourly profit at location j from HVs is $(1-\gamma)P_H \cdot \tau_j \cdot \lambda_{H,j}$. The platform's total profit is given by

$$R_A + (1-\gamma)R_H - C_A N_A \triangleq \sum_{j=1}^{L} (P_A - c_A) \cdot \tau_j \cdot \lambda_{A,j} + (1-\gamma) \sum_{j=1}^{L} P_H \cdot \tau_j \cdot \lambda_{H,j} - C_A N_A,$$

where we define $R_A \triangleq \sum_{j=1}^{L} (P_A - c_A) \cdot \tau_j \cdot \lambda_{A,j}$ and $R_H \triangleq \sum_{j=1}^{L} P_H \cdot \tau_j \cdot \lambda_{H,j}$. We also assume that using AVs is marginally more profitable for the platform than using HVs (i.e., $P_A - c_A > (1-\gamma)P_H$).

Human wage equilibrium. The number of HVs, N_H, is decided by a wage equilibrium (see e.g., Hall et al. (2021)). In particular, the expected total earnings of HVs must be equal to their total opportunity cost at equilibrium:

$$\gamma \cdot R_H = \gamma \cdot \sum_{j=1}^{L} P_H \cdot \tau_j \cdot \lambda_{H,j} = r \cdot N_H, \tag{3.4}$$

where $r \in (0, P_H)$ denotes the corresponding reserve earnings of drivers, i.e., what they could make outside the system. If $\gamma R_H < r N_H$, some drivers would leave the system, and N_H would decrease until $\gamma R_H \geqslant r N_H$, or $N_H = 0$. Conversely, if $\gamma R_H > r N_H$, more human

[3] In practice, Uber takes a 25% commission ($\gamma = 0.75$): `https://www.uber.com/gh/en/drive/basics/tracking-your-earnings/`, last accessed: 2022-12-02

[4] In the queueing model, we focus on operational controls and assume that the prices are exogenously chosen such that riders are indifferent between AVs and HVs. However, we will relax this assumption in Section 3.6 and show that the main results hold when the platform controls the pricing policy to affect riders' preferences.

drivers would be willing to join the platform, and N_H would increase until $\gamma R_H \leqslant rN_H$. Therefore, at equilibrium, N_H must verify $\gamma R_H = rN_H$.

Platform's problem. The platform chooses a dispatch policy π and balances the vehicles to maximize its hourly profit given the average number N_A of AVs available and anticipating the human driver equilibrium. Because we focus on how the service level is impacted as more AVs are introduced to the market, the number of AVs, N_A, is assumed to be exogenously given and C_A is omitted in most of our results except for Proposition 9 where we discuss the optimal N_A. The platform's main problem is

$$
\begin{aligned}
\sup_{\substack{\pi\in\Pi, N_{i,j}, \lambda_{i,j}, \\ \forall i\in\{A,H\}, j\in\{1,\dots,L\}, N_H}} \quad & \sum_{j=1}^{L} (P_A - c_A)\tau_j \cdot \lambda_{A,j} + (1-\gamma) \sum_{j=1}^{L} P_H \tau_j \cdot \lambda_{H,j} \\
\text{s.t.} \quad & \sum_{j=1}^{L} N_{i,j} = N_i, \quad i \in \{A, H\}, \\
& (\tau_j + W^{\pi}_{i,j}(\lambda_{A,j}, \lambda_{H,j})) \cdot \lambda_{i,j} = N_{i,j}, \quad i \in \{A, H\},\ j \in \{1, \dots, L\}, \\
& \gamma \cdot \sum_{j=1}^{L} P_H \cdot \tau_j \cdot \lambda_{H,j} = r \cdot N_H, \\
& \lambda_{A,j} + \lambda_{H,j} \in [0, \mu_j), \qquad j \in \{1, \dots, L\}.
\end{aligned}
\tag{$\mathcal{M}$}
$$

One of the main challenges in analyzing Problem ($\mathcal{M}$) and deriving insights from its solution is that the space of policies is general, and, therefore, we do not have available closed-form expressions for the waiting times. To tackle this challenge, in Section 3.3, we use the achievable region approach (Bertsimas (1995), Dacre et al. (1999)) to obtain a reformulation of Problem ($\mathcal{M}$). This allows us to establish a sharp characterization of the per-location and city-wide service levels.

3.3 Solution to the Queueing Model

In this section, we develop a tractable approach to solve $(\mathcal{M})$. The first step is to reduce the space of feasible policies to the space of non-idling policies.

Lemma 2 (Optimality of non-idling policies)**.** The optimal dispatch policy in $(\mathcal{M})$ is non-idling.

All the proofs of the study are available in the appendix. Intuitively, rejecting requests leads to delays in the form of longer idle times for vehicles and, consequently, to lower platform profit.

Lemma 2 is an intuitive but necessary result: as vehicles are always dispatched when a request arrives, our model coincides with that of a matching queue setting in which, at each location j, drivers wait to be matched to requests that arrive at rate μ_j. In turn, the total waiting time of drivers at location j, $W_j^\pi(\lambda_{A,j}, \lambda_{H,j})$, satisfies

$$W_j^\pi(\lambda_{A,j}, \lambda_{H,j}) = \frac{1}{\mu_j - \lambda_{A,j} - \lambda_{H,j}}. \tag{3.5}$$

This expression holds for any non-idling policy and is what makes our model tractable. It can be obtained from the traditional M/M/1 average wait time, as we can interpret the next vehicle to be dispatched as "being served" and ignore the vehicle type.

We now demonstrate how to transform Problem $(\mathcal{M})$ into a simplified form that can enable theoretical analysis. We use the "achievable region approach" from the stochastic optimization literature (see, e.g., Bertsimas (1995) and Dacre et al. (1999)). In this approach, instead of directly solving over the space of policies Π, we optimize over an alternative space of a judiciously chosen metric that varies with the policy choice. More precisely, we choose the total throughput rate of each location j, $\lambda_j = \lambda_{A,j} + \lambda_{H,j}$ as the metric and provide a reformulation of $(\mathcal{M})$ that uses $\{\lambda_j\}_{j=1}^L$ as variables instead of $\boldsymbol{\lambda}_A$ and $\boldsymbol{\lambda}_H$.

First, we note that the objective in Problem $(\mathcal{M})$ can be cast purely in terms of $\{\lambda_j\}_{j=1}^L$. In fact, from Equation (3.5), we have that the average waiting time at each location j across

types is $W_j = 1/(\mu_j - \lambda_j)$. In turn, Little's law and the wage equilibrium constraint translate into

$$\sum_{j=1}^{L} \left(\tau_j + \frac{1}{\mu_j - \lambda_j} \right) \cdot \lambda_j = N_A + N_H \quad \text{and} \quad rN_H = \gamma P_H \cdot \sum_{j=1}^{L} \tau_j \cdot (\lambda_j - \lambda_{A,j}).$$

These identities allow us to reformulate the objective of $(\mathcal{M})$ in terms of $\{\lambda_j\}_{j=1}^{L}$:

$$\sum_{j=1}^{L} (p - \hat{r}) \cdot \tau_j \cdot \lambda_j + \hat{r} \cdot N_A - \hat{r} \cdot \sum_{j=1}^{L} \frac{\lambda_j}{\mu_j - \lambda_j}.$$

where $p \triangleq P_A - c_A$ and $\hat{r} \triangleq r[P_A - c_A - (1-\gamma)P_H]/(\gamma P_H)$.

Second, note that given the above, if we can fully characterize the achievable values of $\{\lambda_j\}_{j=1}^{L}$, we would be able to solve problem $(\mathcal{M})$ in the space of rates rather than in the space of policies (contingent on being able to find a policy that implements the optimal rates). We can accomplish this by exploiting Little's law and the wage equilibrium. Indeed, we establish that the achievable values of $\{\lambda_j\}_{j=1}^{L}$ lie in a bounded polyhedron of $\mathbb{R}^L$ that has extreme points that come from two natural policies in Π: one that always matches requests to AVs first—it fully prioritizes AVs—and another that always matches requests to HVs first—it fully prioritizes HVs. Among non-idling policies, full prioritization of AVs leads to the lowest total number of vehicles because it creates fewer incentives for HVs to join the platform by giving an advantage to AVs. In contrast, by fully prioritizing HVs, we provide maximal incentives for this type of vehicle to join, and hence, the total number of vehicles in the system is maximized. Hence, the total number of vehicles in service, $\sum_{j=1}^{L} \tau_j \cdot \lambda_j$, from fully prioritizing AVs is lower than the total number of vehicles in service from fully prioritizing HVs, and these are the lowest and largest possible total number of vehicles in service. Letting $\lambda_j^{\dagger}$ and $\lambda_j^{\ddagger}$ be the throughput rate at location j from fully prioritizing AVs and HVs respectively,[5] the following proposition summarizes this discussion.

[5]See Appendix B.1.2 for the precise definition of $\lambda_j^{\dagger}$ and $\lambda_j^{\ddagger}$.

Proposition 6 (Achievable Region)**.** Let $\boldsymbol{\lambda}$ be an optimal solution to $(\mathcal{M})$ then

$$\boldsymbol{\lambda} \in \mathcal{A} \triangleq \left\{ \{\lambda_j\}_{j=1}^L \in \mathbb{R}_+^L : \sum_{j=1}^L \tau_j \lambda_j^\dagger \leq \sum_{j=1}^L \tau_j \cdot \lambda_j \leq \sum_{j=1}^L \tau_j \lambda_j^\ddagger \right\}.$$

Proposition 6 leads to the following reformulation of Problem $(\mathcal{M})$ in which we now optimize over the rates $\{\lambda_j\}_{j=1}^L$, given by

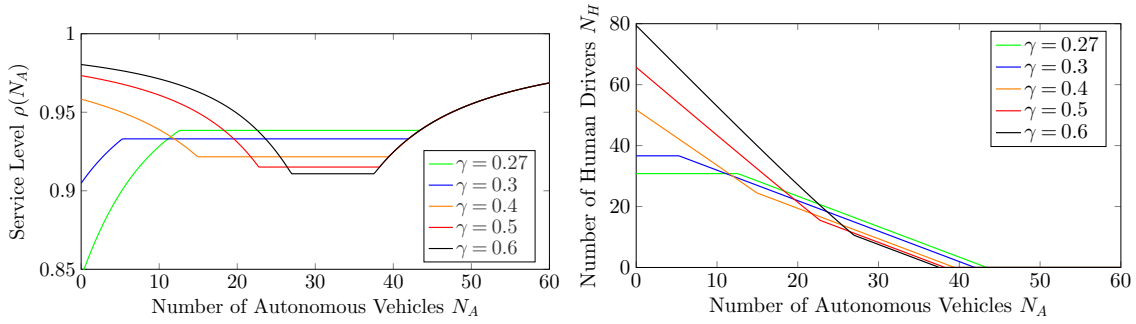
$$\begin{aligned} \max_{\{\lambda_j\}_{j=1}^L} \; & \sum_{j=1}^L (p - \hat{r})\tau_j \lambda_j + \hat{r} N_A - \hat{r} \sum_{j=1}^L \frac{\lambda_j}{\mu_j - \lambda_j} \\ \text{s.t.} \quad & \{\lambda_j\}_{j=1}^L \in \mathcal{A}, \\ & \lambda_j \in [0, \mu_j), \qquad\qquad j \in \{1, \ldots, L\}. \end{aligned} \tag{$\mathcal{M}'$}$$

Problem $(\mathcal{M}')$ represents an upper bound of Problem $(\mathcal{M})$. However, in principle, the feasible region may contain rates that do not correspond to a dispatch policy. We show that the latter is not true. When the rates are such that total revenue hits one of the boundaries in $\mathcal{A}$, the corresponding dispatch policy will be to prioritize either AVs or HVs fully. When the rates are in the interior of $\mathcal{A}$, then the corresponding dispatch policy will partially prioritize vehicles; for example, each request will be prioritized towards AVs with a certain probability. Hence, the optimal solution to Problem $(\mathcal{M}')$ can be implemented in the original Problem $(\mathcal{M})$. The following proposition formalizes this.

Proposition 7 (Reformulation)**.** Problem $(\mathcal{M})$ is equivalent to Problem $(\mathcal{M}')$.

3.4 One Location: Impact of AVs on Service Level

We now use the tractable formulation of the previous section to derive insights. At this point, we focus on a model with $L = 1$ location to build intuition and understand the impact of the AV supply N_A and the pay ratio γ on the service level. We will switch to $L > 1$ in the following section to study the spatial effects. Therefore, we omit the location subscript in this section.



(a) (b)

Figure 3.3: Service level and number of HVs for the optimal policy. We use $L = 1$, $\mu = 30$, $P_H = 9, P_A = 8$ and $c_A = 0.5, r = 2, \tau = 1$.

We start with a study of a numerical solution to an instance of Problem ($\mathcal{M}'$), shown in Figure 3.3. We then present formal results that support our observations. The left panel of the figure shows the service level change as we vary N_A and γ. When γ is high enough (here, $\gamma \geq 0.4$), we identify three phases as the number of AVs increases: the service level first decreases, then stabilizes, and then increases again. For smaller values of N_A, the platform uses a dispatch policy that fully prioritizes AVs, which reduces the service level. Indeed, as shown in the right panel of Figure 3.3, prioritizing AVs lowers the earnings of human drivers and drives them out of the market at a higher rate than AVs enter the market: fewer cars are available, which lowers the service level. However, reducing the number of HVs reaches a breakeven point, where the hit on revenue due to the lack of HVs is higher than the cost advantage of prioritizing AVs. At this point, the platform reduces the AV prioritization, leading to the flat part of the curves in the left panel of Figure 3.3. Note that the platform "partially prioritizes" the AVs less and less as N_A increases in a way that maintains a constant service level (the flat part of each curve in the figure), which decreases in γ. Interestingly this optimal partial prioritization is set so that each added AV substitutes exactly one HV in equilibrium. However, when the number of AVs is sufficiently high, the HVs are entirely driven out of the market. In this third phase, the service level increases

with N_A because all HVs are gone, and the total number of vehicles grows. We summarize these observations in the following theorem.

Theorem 4 (AVs can reduce the service level)**.** *If γ is large enough and there are HVs in the market at equilibrium, then as N_A increases, the service level decays until the number of AVs is sufficiently high. After that, the service level stays constant and then increases when the number of AVs reaches a threshold. At that point, HVs are completely driven out of the market.*

To better understand this result, we can use Little's law to relate the service level to the total number of vehicles:

$$N_A + N_H = W^{\pi}(\lambda_A, \lambda_H) \cdot (\lambda_A + \lambda_H) = \frac{\lambda_A + \lambda_H}{\mu - \lambda_A - \lambda_H} = \frac{\rho}{1 - \rho},$$

where $\rho = (\lambda_A + \lambda_H)/\mu$ is the service level in the system. Therefore, the service level is simply an increasing function of the total number of vehicles. In turn, the total number of vehicles is a function of the platform's prioritization policy. If the platform prioritizes AVs, it will "starve" the HVs. Any additional AV supply will quickly reduce the number of HVs, potentially far more than one HV for each AV, reducing the total supply and, therefore, the service level. This happens up to the point where this supply loss is too costly, and the platform will then reduce the AV prioritization to maintain the number of vehicles. The figure shows the driving force behind the reduction in service levels. Fully prioritizing AVs expels HVs quickly. The higher γ is (costly HVs), the faster the optimal policy expels HVs.

Such an AV prioritization policy is only helpful if γ is not too low. When γ is low ($\gamma < 0.4$ in Figure 3.3), an entirely different strategy is optimal. Indeed, when γ is low (HVs are not paid much), the trips served with HVs are more profitable for the platform than when γ takes a higher value. Even a slight deprioritization of HVs may lead to a massive loss due to HVs leaving the market in equilibrium. Hence, even if HV trips are more costly than AV trips, it is still preferable to avoid reducing the requests sent to HVs. Accordingly, fully prioritizing HVs guarantees that we are not losing HVs when adding AVs (they are not affected by the

extra supply), which increases the total number of vehicles available and, hence, the service level. We formalize this insight in the following proposition:

Proposition 8 (Prioritizing HVs can increase the service level)**.** When γ and N_A are sufficiently small, it is optimal to prioritize HVs, and the service level increases in N_A.

Prioritizing HVs may seem counter-intuitive because if the platform were myopic and considered that N_H was constant, it would prioritize AVs as they are cheaper. As a consequence, Proposition 8 is the result of the platform anticipating the equilibrium behavior of the drivers. However, we expect this situation to be rare in practice. Indeed, a low pay ratio γ is not the norm, as it would mean that the platform does not have enough drivers and has a much lower service level (see the left panel of Figure 3.3 when $N_A = 0$), which would not be practical in a competitive setting.

Finally, suppose the platform can decide the supply of N_A when the capital cost of AVs is C_A per vehicle-hour. That is, we modify problem $(\mathcal{M})$ so that N_A is now a decision variable of the platform, and we add a cost $-C_A \cdot N_A$ to the objective function. The following result describes the optimal solution to this optimization problem:

Proposition 9 (Optimality of operating AVs only)**.** Under the optimal choice of N_A, HVs are entirely replaced with AVs (e.g., $N_H = 0$) if and only if $C_A \leq \hat{r}$. If $C_A > \hat{r}$, the platform may operate both AVs and HVs (i.e., it is possible to have $N_A, N_H > 0$).

As mentioned in Section 3.1, the initial launch of commercial AVs may be limited, expensive, and risky, so the cost of AVs might be high (e.g., $C_A > \hat{r}$) in the early days of AV deployment (Litman, 2023). Even in that case, Proposition 9 states that it could be optimal to operate a mixed fleet with $N_A > 0$, motivating the research question of this work. If AVs become cheaper than HVs, however, using human drivers is not profitable. In Figure 3.4, we illustrate Proposition 9 and show the profits that the platform can earn for each value of N_A and γ, where $N_A^\star(\gamma)$ is the optimal number of AVs given γ. Compared with the right panel in Figure 3.3, we can see that in this example, $N_A^\star(\gamma)$ is always less than the threshold

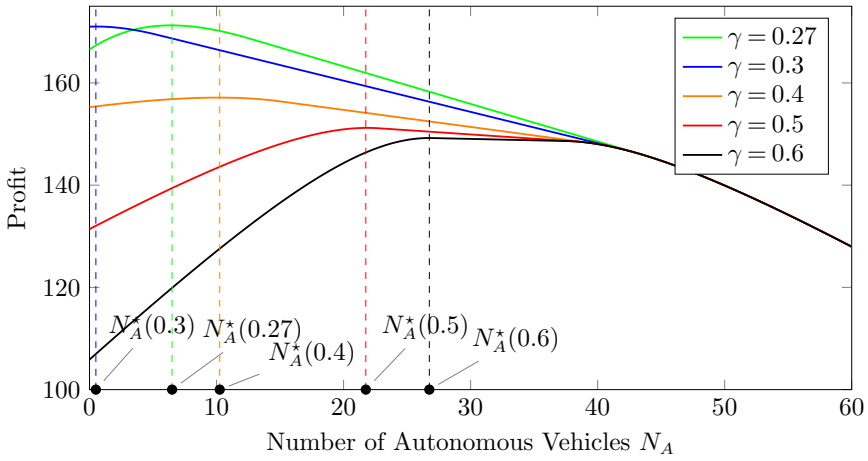


Figure 3.4: Profit for the optimal policies with respect to N_A. We use $L = 1$, $\mu = 30$, $P_H = 9, P_A = 8$ and $c_A = 0.5, C_A = 1.5, r = 2, \tau = 1$. $N_A^\star(\gamma)$ is the optimal number of AVs given the value of γ.

of N_A such that all the HVs leave the market, confirming that a hybrid solution using both AVs and HVs can be optimal when $C_A > \hat{r}$.

Intuitively, although the cost of AVs is higher than $\hat{r}$, the platform is still able to use the AVs more efficiently than the HVs thanks to the use of prioritization. There are two cases here: for high values of γ, it can be profitable to introduce AVs and prioritize them, whereas, for low values of γ, it can be profitable to introduce AVs while still prioritizing HVs. Remember from the human driver equilibrium (Equation (3.4)) that, for fixed r and HV revenue, the HV supply N_H is an increasing function of γ — drivers are paid more, so there are more drivers. If γ is high, N_H may be "too high" compared to what is really needed. Therefore, we can invest in expensive AVs and prioritize them so that they serve as many rides as possible: even if $C_A > \hat{r}$, each AV-hour can be much more productive than HV-hours, justifying the investment. This is possible because N_H is too high, and we can withstand the loss of human drivers due to AV prioritization. However, as γ decreases, this is not the case anymore, and N_A^* decreases correspondingly. When γ is particularly small, the platform will start to prioritize HVs: in that case, N_H is too small in equilibrium (we

are not paying them enough), so we want to prioritize the HVs to preserve the few cheap HVs. However, it is still worth introducing the more expensive AVs because there are still a lot of profitable rides left to serve, as we can see $N_A^\star(0.27) > N_A^\star(0.3)$.

3.5 Multiple Locations: Spatial Inequality

In the previous section, we established that introducing AVs can adversely impact the system-wide service level (despite increasing the platform's profit). In this subsection, we explore the spatial dimension of this effect. In particular, we are interested in understanding how the service levels in different regions in a city are affected by the introduction of AVs.

Consider our initial setting with multiple locations and, without loss of generality, suppose that $\mu_j \tau_j \geqslant \mu_{j+1}\tau_{j+1}$ for $j \in \{1, \ldots, L-1\}$. Our model can illustrate a region with a range of demand density, from urban areas with high demand ($j = 1$) to rural areas with low demand ($j = L$).[6] Let $\rho_j^\star(N_A)$ be the service level of the optimal solution to Problem $(\mathcal{M})$ at location j for a given N_A. The following result establishes a form of spatial inequality across locations.

Theorem 5 (Spatial inequality of the service level)**.** *In any optimal solution of Problem* $(\mathcal{M})$*, higher-demand locations have a higher service level:* $\forall j < L$, $\rho_j^\star(N_A) \geqslant \rho_{j+1}^\star(N_A)$*).*

Additionally, out of all locations j *that have positive supply in the optimal solution* ($N_{A,j} + N_{H,j} > 0$*), the change of service level with respect to* N_A *is larger in low-demand locations (i.e.,* $|\frac{\partial \rho_j^\star}{\partial N_A}| \leqslant |\frac{\partial \rho_{j+1}^\star}{\partial N_A}|$*).*

Theorem 5 establishes that different locations in a city are affected differently by the introduction of AVs. The first part of the theorem is not the most surprising: no matter whether HVs are present in the optimal solution, locations with high demand have better

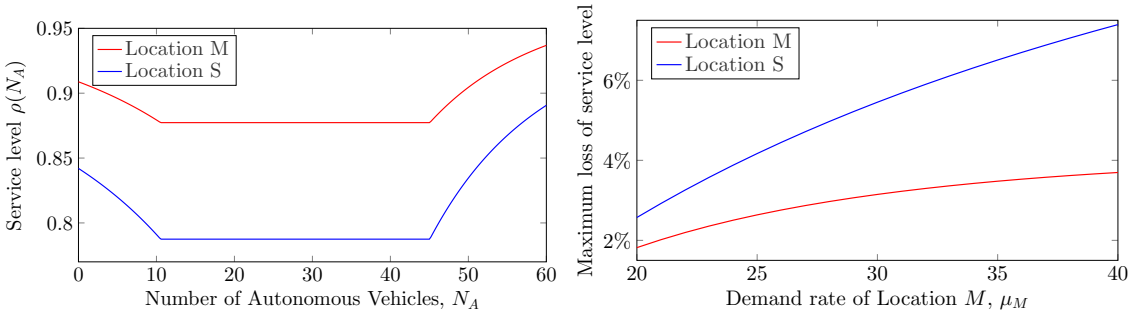
[6] In 2019, 45% of the urban residents in the US have used a ride-hailing app while only 19% of Americans living in rural zones have done so (Jiang, 2019).

service levels than locations with low demand. Intuitively, high-demand areas are more profitable than low-demand areas because of economies of scale: we can achieve the same service level with a higher vehicle utilization in a high-demand area. Hence, the platform is better off serving more trips there regardless of the number of AVs. But the second part of Theorem 4 introduces a worse and more novel effect. Adding AVs to a given city will *worsen* the spatial inequalities, affecting the low-demand locations more than the high-demand locations. Essentially, the platform will choose to concentrate the AVs into the more profitable areas to maximize profit. But introducing AVs will make HVs leave, mainly at the expense of the less profitable areas, which will lose more drivers and end up experiencing the most significant reduction of service level. The following result sheds light on the driving force behind this.

Proposition 10 (Imbalance of driver concentration)**.** When γ is high enough, and N_A is low enough, the platform fully prioritizes AVs across locations. When this happens, as we add AVs, they will concentrate on high-demand locations, and HVs will leave higher-demand locations more than the low-demand ones. That is, $\frac{\partial N_{A,j}}{\partial N_A} \geqslant \frac{\partial N_{A,j+1}}{\partial N_A} \geqslant 0$ and $\frac{\partial N_{H,j}}{\partial N_A} \leqslant \frac{\partial N_{H,j+1}}{\partial N_A} \leqslant 0$.

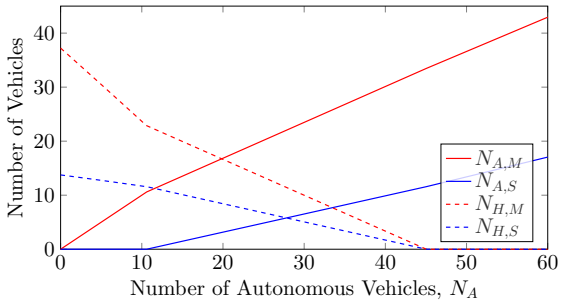
This is the multi-location extension to Theorem 4. Because γ is high, the platform benefits from prioritizing AVs when introducing them. But in the case of spatial heterogeneity, the platform will concentrate them in high-demand locations where they can get the highest utilization: $\frac{\partial N_{A,j}}{\partial N_A} \geqslant \frac{\partial N_{A,j+1}}{\partial N_A}$. Therefore, there is not a lot of demand left for humans to serve in high-demand locations, and they are relocated to low-demand ones, $\frac{\partial N_{H,j}}{\partial N_A} \leqslant \frac{\partial N_{H,j+1}}{\partial N_A}$. Overall, the total number of vehicles decreases, and so does the service level. However, the increase of AVs in high-demand areas is enough to lead to a lower service level reduction compared to low-demand areas. Therefore, Proposition 10 reveals another form of spatial inequality. The introduction of AVs affects not only the service quality but also the distribution of the types of vehicles. To increase the utilization of AVs and maximize profit, the platform prefers to deploy AVs in busier downtown areas. At the same time, human drivers

have to concentrate on the outskirts or other less dense areas.



(a) Service level between M and S under the optimal policy.

(b) The maximum loss of service levels with respect to μ_M.



(c) Number of AVs and HVs at each location.

Figure 3.5: An example with two locations M and S, where location M has a more demand than location S. We use $P_A = 1.9, P_H = 2, c_A = 0.1, r = 0.7, \tau_M = \tau_S = 1$, and $\mu_M = 30, \mu_S = 10, \gamma = 0.5$.

In Figure 3.5, we consider an example with two locations M and S, where location M has more demand than location S (i.e., $\mu_M \tau_M > \mu_S \tau_S$). Figure 3.5a is the multi-location version of Figure 3.3: it illustrates Theorem 5 and shows that the low-demand location has a lower service level and is more negatively affected by the introduction of AVs. Figure 3.5b depicts the maximum loss of service level for various values of μ_M, which is the difference between the service level when there are only HVs in the market (i.e., $N_A = 0$) and when the service level reaches the minimum point for some value of N_A. We can see that the maximum losses of the service level at the two locations M and S increase in μ_M, even if we do not change

the demand at this location. This is another form of unfairness and coupling between the two locations. The decrease in service level at the low-demand location is more substantial if the demand at the high-demand location becomes larger. As the difference in hourly trips between M and S increases, the platform finds it profitable to dedicate even more of the M demand to AVs at the expense of the service level of location S. The following proposition confirms this result.

Proposition 11 (Maximum loss of service level)**.** Suppose that i, j are two locations such that $N_{A,j} + N_{H,j} > 0$ (e.g., positive supply) for any N_A, and that $i < j$ (i has higher demand). Then, the maximum loss of service level is larger at j. Additionally, as $\mu_i \tau_i \to \infty$, the maximum loss at j converges to $\sqrt{\frac{\hat{r}}{\mu_j \tau_j (p-\hat{r})}}$, while the maximum loss at i converges to zero.

All in all, the service level in low-demand locations can be severely affected compared to high-demand locations.

Finally, Figure 3.5c shows the number of different vehicles across locations. As AVs enter the market, they are allocated to the high-demand location, which forces some HVs to leave both the high and low-demand locations. When N_A exceeds some threshold, AVs start to enter S, and the decline rate of N_H reduces at M but rises at S. The growth rate of N_A and the exit rate of N_H are always larger in location M than in location S, when AVs are fully prioritized. This further showcases the spatial inequality effect: the platform prefers to operate high-demand locations with AVs, which, in turn, relegates HVs to the low-demand, less profitable, locations.

3.6 Simulation Study

While our queueing model is chosen to be tractable and illustrate important effects, it does not account for some realistic factors in a ride-sharing market, such as pricing, pick-up times, and vehicle relocation. In particular, our model assumed that the platform could "assign"

any vehicle type to any customer, and we did not consider the customer preferences and sensitivity to prices, waiting time, and vehicle type. To confirm our queueing model findings in settings that do account for these factors, we develop a ride-hailing simulator using New York City (NYC) data. The simulation study not only serves as a robustness check for our theoretical results of the queueing model, but it also reveals additional, more granular, spatial effects of introducing AVs on service levels.

3.6.1 Simulation Description

We now describe the dataset and the simulation design, emphasizing the extra features that are missing in the queueing model. Some details are omitted for the sake of brevity, a complete description is available in Appendix B.3 in the appendix.

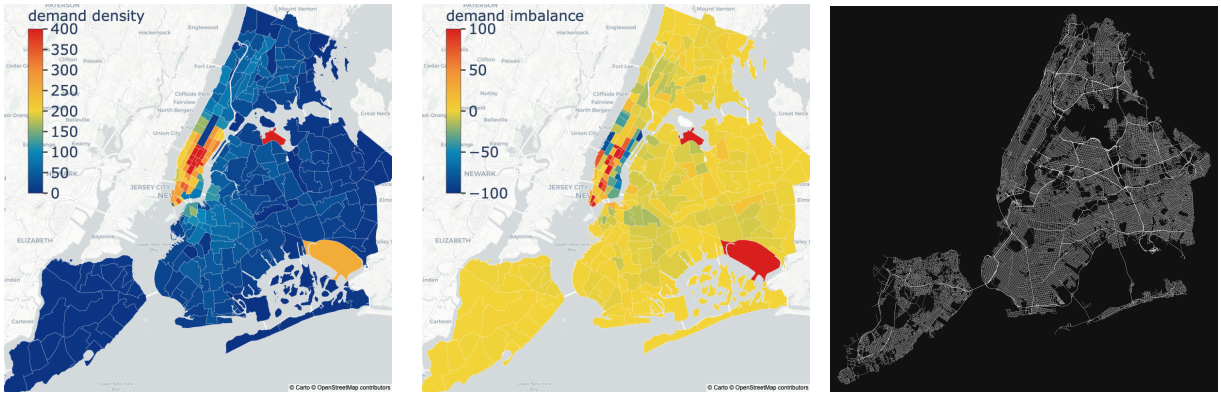
3.6.1.1 Data Processing

To estimate the parameters in our model, we use the New York City Open Data platform to access the historical record of High-Volume For-Hire Vehicle (HVFHV) data.[7] For each trip in NYC, this data contains the origin, destination, and request time stamp for Lyft, Uber, and Via (the three leading ride-hailing platforms in NYC in 2020). For a more balanced demand distribution across the city (inflow and outflow are similar within zones), we consider the trip data between 11 AM and 1 PM during workdays in January 2020 (the month before the coronavirus led to a significant drop in demand), which corresponds to a total of 1,093,431 recorded trips.[8] For privacy reasons, the exact origin and destination of each trip are unavailable, but we have access to their "taxi zones." NYC is divided into 257 such zones, as illustrated in Figure 3.6. To generate exact locations and travel time, we use

[7] `https://data.cityofnewyork.us/Transportation/2020-High-Volume-FHV-Trip-Records/yrt9-58g8`, last accessed: 2022-12-08.

[8] In Section 3.6.1.2, we will discuss how demand imbalance may influence our results.

OpenStreetMap[9] to obtain the road network of NYC, which can be visualized in Figure 3.6c. We assume that each trip starts and ends in a uniformly random intersection from the origin and destination zone, respectively (c.f. Appendix B.3.1). Figure 3.6a illustrates the demand density, which is obtained by dividing the hourly trips of a zone by its area,[10]. We assume that the driver pay ratio is $\gamma = 75\%$ and we set their reserve earnings to be $r = \$33$ per hour, which is derived from NYC data as discussed in Appendix B.3 in the appendix.



(a) Demand density (hourly trips per km^2)

(b) Demand imbalance (hourly trips per km^2)

(c) NYC street network

Figure 3.6: Inputs to the fleet balancing problem in NYC, for each zone.

3.6.1.2 Simulation

Our simulation shares many features with the queuing model, such as driver earning equilibrium, commission rate, and a profit-maximizing firm. However, a key distinctive feature is the incorporation of the dynamic vehicle evolution within a realistic spatial environment. For each customer taking a ride, we assign a vehicle to pick them up and transport them along the network to their destination. Subsequently, the vehicle may either remain idle,

[9] `https://osmnx.readthedocs.io/en/stable/`, last accessed: 2022-12-08.

[10] For the airports, we used an estimate of the "driving area" of 1km^2 rather than counting the full airport area, because the driving area in an airport is usually much smaller than its actual area

serve another customer, or reposition. In addition to integrating the spatial dynamics of a ride-hailing company's operations, we enrich our model by considering customer sensitivity to pricing, pickup times and vehicle type, pricing strategies, and vehicle relocation. Next, we discuss how we incorporate these aspects in the simulator and then its limitations.

Customer's decision. Mirroring the typical interface of ridesharing platforms providing multiple transportation options, we assume that a customer facing a decision to travel has three options: selecting the nearest available HV, selecting the nearest available AV, or canceling the trip altogether. This choice is determined by a utility model that takes into account the vehicle type, hourly pricing, travel time from the origin to the destination, and the estimated time of arrival (ETA). Specifically, for a customer i requesting a trip, the customer's utility $U_{i,j}$ when assigned a vehicle j is given by:

$$U_{i,j} \triangleq [a_0 + \theta \mathbf{1}_{\text{j is an AV}} - (P_{A,i}\mathbf{1}_{\text{j is an AV}} + P_{H,i}\mathbf{1}_{\text{j is an HV}})] \times \text{travel time}_i + a_1 \text{ETA}_{i,j}$$

where $a_0 > 0$ represents the utility of being transported by an HV (per unit of time), θ is a correction term if the vehicle is an AV. For example, a positive θ represents the fact that passengers prefer AVs to HVs, everything else being equal., $P_{A,i}$ and $P_{H,i}$ are the price rates (to be chosen by the platform) shown for customer i for AVs and HVs, respectively. Therefore, if AVs are more cost-efficient for the platform, it can choose $P_{A,i} < P_{H,i}$ to incentivize the passengers to choose this option — the AV prioritization can be achieved through pricing rather than the direct operational prioritization of the queueing model. Finally, $a_1 < 0$ models the customer's sensitivity to waiting time before pickup: for example, even if customers prefer AVs and the AV price is advantageous, they could still prefer to request an HV if the closest one is much closer than the AV counterpart.

The utility $U_{i,j}$ measures a consumer's surplus from riding vehicle j in dollars. If the utility for any available vehicle j is negative, the consumer will cancel the request. If at least one available vehicle j yields a positive $U_{i,j}$, the customer will choose the option that maximizes her utility, which is equivalent to choosing the nearest available AV or HV after considering pricing and vehicle preferences. The selected vehicle then proceeds to the pickup

location and transports the customer to her destination. Upon completing the trip, the vehicle will wait at the destination for the next request or follow a relocation mechanism that we will describe later.

In our baseline simulation setting, we assume that the customers are indifferent between AVs and HVs (i.e., $\theta = \$0$/hour), and that the operational cost of AVs is negligible (i.e., $c_A = \$0$/hour). In addition, we set $a_0 = \$80$/hour, $a_1 = -\$20$/hour. These numbers are chosen such that, in an HV-only market at equilibrium, a firm choosing a uniform price rate would choose $P_{H,i} = \$75$/hour to maximize profit, resulting in an overall service level (the fraction of fulfilled requests in NYC) of about $\rho = 90\%$. We will explore the robustness of our results by varying these parameters in Section 3.6.3.

Pricing and Prioritization. In contrast to the queuing model, in the simulation, customers can choose the type of vehicle they want to take, and the platform can only influence customers' choices through a pricing policy. Due to the complexity of the system (mixed fleet and extremely large state space), we limit ourselves to a manageable set of pricing policies that nonetheless enable the platform to prioritize AVs or HVs and to affect the respective distributions of AVs and HVs in the city. For example, our set of policies allows the platform to potentially keep the AVs in the high-demand areas (downtown) by preventing the use of AVs for rides that lead to low-demand areas. Our main limitation is that we need to run thousands of highly detailed simulations to evaluate HV equilibria, and we chose the "richest" set of policies that we could still optimize over within less than one month of computation on a university cluster.

We let the price rate $P_{A,i}$ (or $P_{H,i}$) offered to customer i to be equal to a base price that is identical for all the trips, plus an adjustment depending on whether the destination of a trip is Manhattan (i.e., the area with the highest demand). That is, for any customer i, the price rates are

$$P_{type,i} = P_{type,base} + \delta_{type}(2 \times \mathbf{1}_{\text{destination of i is in Manhattan}} - 1) \quad type \in \{A, H\}$$

where $P_{type,base}$ is the base price of vehicle-type "*type*", δ_{type} is the destination-based adjustment, and $\mathbf{1}_{\text{destination of i is in Manhattan}}$ is an indicator function.

This four-parameter $(P_{A,base}, P_{H,base}, \delta_A, \delta_H)$ class of pricing policies is rich enough to give the platform the flexibility to both prioritize and allocate (to a particular region) any specific type of vehicle. For example, choosing $P_{A,base} < P_{H,base}$ allows the platform to increase the utilization of AVs by offering a discount. Additionally, choosing $\delta_A < 0$ means that the AV price is increased if the request destination is outside Manhattan and decreased if it is inside Manhattan. This means that the platform is trying to concentrate the AVs in the higher-demand areas, incentivizing customers towards AVs if they want to go to or stay within Manhattan. We solve for a pricing policy through a two-stage grid search: we first find the best (profit-wise) $P_{A,base}$ and $P_{H,base}$ assuming $\delta_A = \delta_H = 0$ then fix these base price rates and optimize δ_A and δ_H in the second search. This is computationally intensive: for example, given a value of $P_{A,base}, P_{H,base}$, we run a grid of full platform simulations for various values of N_H (the number of human drivers). We find the N_H that satisfies the human driver earning equilibrium (with an average hourly earning r) and evaluate the platform profit for this one particular simulation. We then repeat this process on a grid of values $P_{A,base}, P_{H,base}$ to find the pricing policy that maximizes the equilibrium platform profit. We then repeat this entire process to learn the destination pricing corrections δ_A, δ_H.

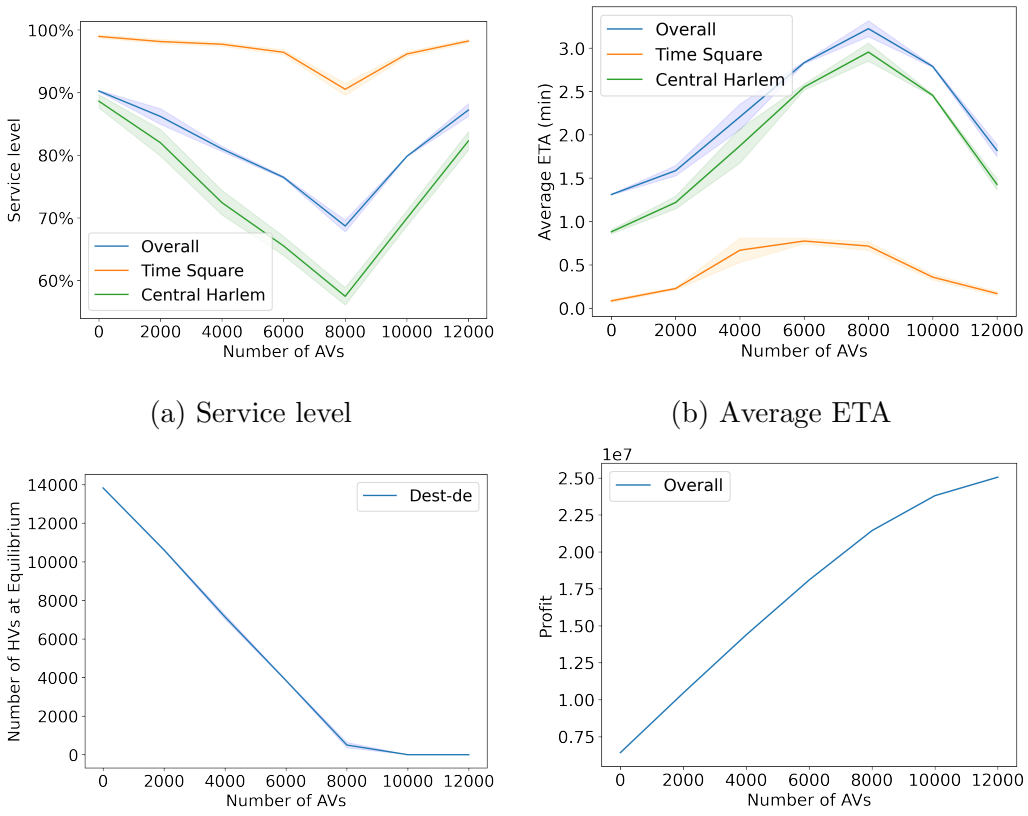
Relocation. Without relocation, most drivers would end up in the same area given enough simulation time, potentially leading to large driver imbalances in the city. To prevent this, we consider a simple relocation policy that periodically relocates vehicles to maintain a balanced distribution. Each vehicle has an exponential clock (with a mean of 2 hours). Once the clock's time is up, and the vehicle is idle, the vehicle is (instantaneously) relocated to a zone that is sampled according to the demand distribution. That is, the probability of choosing the zone is its hourly trips divided by the total hourly trips in the city, and, therefore, vehicles are more likely to reposition to high-demand locations (Afèche et al., 2023). Then the vehicle will be relocated to one of the nodes within the zone, which is uniformly selected

at random. After relocation, a new clock is generated, and the vehicle will restart its work. This can also be interpreted as drivers leaving the platform after driving for a while and new drivers becoming available in the city. While we used this simplified relocation policy for tractability, we checked the robustness of this approach by implementing a state-of-the-art relocation algorithm in Section 3.6.3.

Limitations. In this simulation study, we account for pricing, customer preferences and demand patterns, and spatial aspects such as pick-up times, travel times, and relocation. Most of our choices were meant to implement a model that is as realistic as possible, given our computation limitations. In each run, the simulation must process approximately one million requests and track the status of about ten thousand vehicles in the NYC traffic network with about forty thousand nodes. In addition, we need many runs to find the equilibrium number of HVs for any choice of pricing policy and to repeat this hundreds of times to obtain the optimal price rates. While these computational limitations restricted the set of pricing and relocation policies we could consider, our approach was to provide the platform with enough differentiated control on the hybrid fleet to be able to observe the potential spatial inequalities profit maximization would lead to. We provide a more detailed description of the simulator iterations and the equilibrium and price rate computation in Appendix B.3.2.

3.6.2 Confirmation of the Queueing Model Insight

The theoretical results and insights from the queueing theoretical model continue to hold in the more realistic simulated setting. Indeed, when AVs are added to the fleet, the platform chooses to prioritize AVs (through attractive pricing) and maintain them in high-demand areas (with destination-based price discounts). As a consequence, HVs leave faster, and the service level decreases in a way that is not homogeneous across the city. We focus on showing the consequence of these policies on the service levels, human drivers, and spatial inequalities.



(a) Service level

(b) Average ETA

(c) Equilibrium Number of HVs

(d) Average profit (without considering the capital cost C_A).

Figure 3.7: Introducing AVs increases profit, but decreases service level and increases ETA, especially in the more remote areas. The number of HVs is decreasing when introducing more AVs. The shaded area around the curves shows the 95% confidence interval (it may be invisible when the interval is quite small).

Figure 3.7a shows the service level in NYC as a function of the number of autonomous vehicles, where the service level is defined as the number of served requests divided by the total number of requests. With 0 AVs, there are 14,000 HVs at equilibrium,[11] and the overall service level is 90.2%, but this number diminishes quickly as more AVs are introduced, forcing human vehicles out of the market (see Figure 3.7c). At 8,000 AVs, almost all the HVs leave, and the overall service level reduces to 68.7%. The decrease in service level disproportionately affects the more remote locations. Times Square loses 8.4% in service level, whereas a more remote zone like Central Harlem loses 31.1% in service level.[12] Figure 3.8a shows a more detailed view of the service level for each zone of New York City. The spatial inequality of service level degradation is striking. As confirmed in Figure 3.1, the service level is degraded by up to 39% in the suburban areas, while the airports and Manhattan maintain a good service level. Notice that some remote zones in Staten Island have a smaller service level degradation because their service levels when $N_A = 0$ are already very low. Their service levels hardly reduce further since they can often find vehicle supply from other zones.

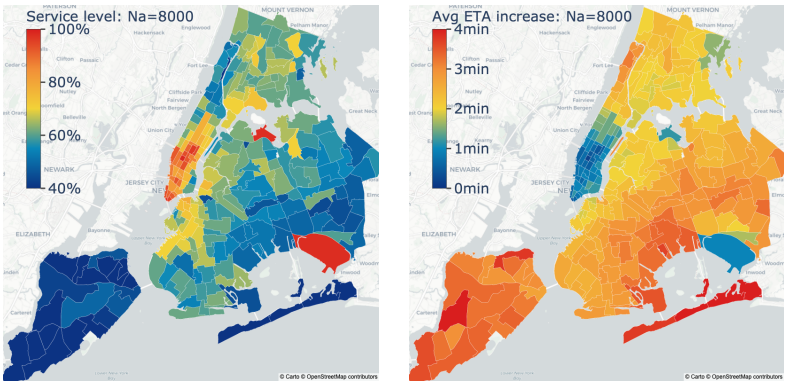
All numbers correspond to the platform's profit-maximizing policy given N_A. In particular, as shown in Figure 3.7d, the platform profit is increasing in N_A despite the negative consequences of the service level. Indeed, more AVs are always preferable (when considering N_A exogenous and therefore not including the AV capital cost C_A), and the degradation in service levels indicates that the platform is simply willing to serve fewer customers in order to use higher-margin AVs more.

Another service quality metric that the simulator enables us to measure is the average time it takes for a vehicle to pick up a customer (i.e., ETA), which we show across NYC in Figure 3.7b. With 0 AVs, the typical wait time is 1.3 minutes,[13] but the average ETA

[11] We explain how an equilibrium N_H is found in Appendix B.3.1 in the appendix.

[12] Note that we selected Times Square and Central Harlem to represent high-demand and low-demand zones, respectively. Historically, Times Square is at the heart of the busiest area in Manhattan, while Central Harlem is an under-served area located at the northern boundary of Manhattan.

[13] These wait times are lower than practice because our travel time estimates are optimistic (see the



(a) The service levels when $N_A = 8000$.

(b) Increase in average ETA from $N_A = 0$ to $N_A = 8000$.

Figure 3.8: Detailed view of the results for each zone in NYC when introducing AVs.

increases quickly as more AVs are introduced. At 8,000 AVs, customers will have to wait about 3.2 minutes on average before a vehicle arrives. Similar to the spatial inequality of the service level, the increase in the average ETA is also disproportionately higher in remote zones, as shown in Figure 3.7b and Figure 3.8b. The average ETA in Times Square only rises by 0.6 minutes, whereas the average ETA can increase by up to 3 minutes in suburban areas.

3.6.3 Robustness Check

We now simulate various scenarios with different parameters, including changes in customer preferences for vehicle types (i.e., θ), the operational cost of AVs (i.e., c_A), and the pay ratio (i.e., γ). Our objective is to assess the impact of these adjustments on our main insights. For each robustness test, we alter only the specific parameter under examination (refer to the first column in Table 3.1) and maintain the other parameters the same as the baseline simulation setting. In Table 3.1, we report the service level with $N_A = 0$ and the service level change of increasing N_A from 0 to $8,000$ in NYC and some specific areas such as Times

appendix).

Square (TS) and Central Harlem (CH). For a more granular view of these results, we have included detailed plots in Appendix B.4.1, offering a visual representation of the robustness checks.

		Service level with $N_A = 0$			Service level change with $N_A = 8,000$		
		NYC	TS	CH	NYC	TS	CH
Customer preferences: θ	60	90.2%	99.0%	88.6%	−24.0%	−12.4%	−33.5%
	10				−19.9%	−7.7%	−30.2%
	0				−21.6%	−8.4%	−31.2%
	−20				−15.0%	−4.3%	−23.7%
	−40				−13.1%	−3.5%	−19.8%
	−60				+4.4%	+1.1%	+8.5%
AV operational cost: c_A	0				−21.6%	−8.4%	−31.2%
	10				−16.5%	−5.3%	−25.2%
	40				−15.8%	−5.4%	−21.7%
	60				+4.0%	+1.0%	+7.4%
Pay ratio:γ	75%	90.2%	99.0%	88.6%	−21.6%	−8.4%	−31.2%
	60%	81.2%	96.7%	72.5%	−10.2%	−5.4%	−13.8%
	50%	46.6%	69.4%	43.7%	+18.3%	+17.2%	+10.2%

Table 3.1: The service levels in the robustness tests with different parameters. The first column indicates the parameter that we change in each robustness test. "NYC" represents the overall New York city, "TS" represents "Time Square", and "CH" represents "Central Harlem".

As we can see, our main insights hold in most of the cases: the introduction of AVs leads to a decline in service levels, with a pronounced disparity in the impact between high-demand (TS) and low-demand areas (CH). This decline is more significant when AVs are more "profitable" (i.e., high γ and low c_A) or more "attractive" to customers (i.e., high θ).

Profitability allows the platform to lower AV prices (P_A), thereby attracting more customers to choose AVs. Similarly, a higher customer preference for AVs results in increased AV usage, even at higher prices. The utilization of AVs is higher in both scenarios, leading to the exit of HVs and the reduction in service levels. However, we also observed some scenarios where the introduction of AVs leads to an increase, rather than a decrease, in service levels. This occurs particularly when AVs are considerably less profitable or attractive compared to HVs. If γ is quite low or c_A is high, HVs are comparatively so cheap that the platform is willing to increase the utilization of HVs instead of AVs. This is consistent with what we found in Figure 3.3 and Proposition 8. Similarly, if customers have a strong aversion to AVs (low θ), reducing P_A cannot effectively increase the utilization of AVs. In both of the cases, the utilization of HVs is higher and the service levels are increased. Nonetheless, as shown in Appendix B.4.2, the spatial inequality proved in Theorem 5 holds in all the cases, as we can see that the service levels change more significantly in low-demand areas than in high-demand areas in general.

We also implement an alternative HV relocation strategy in which we solve a linear program that finds the relocation of vehicles that optimizes the total revenue of HVs as introduced in Braverman et al. (2019).[14] While the alternative relocation method can increase the total revenue of HVs, the service level still declines and the spatial inequality of the degradation still exists. Specifically, the overall service level in NYC drops from 89.8% to 70.3%, with Times Square experiencing a decrease from 94.3% to 90.7%, and Central Harlem from 90.2% to 60.3%.

3.6.4 Other Spatial Effects on Service Level

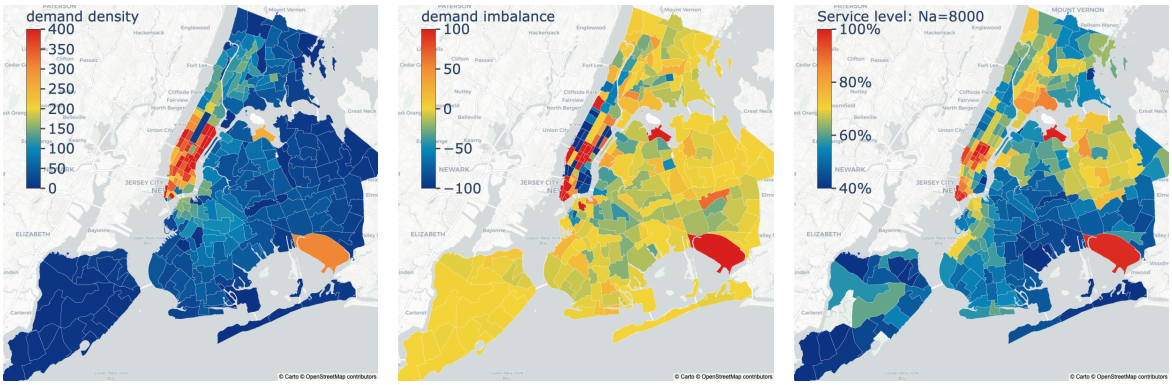
The simulation validates our analysis in Section 3.4 and Section 3.5, but it also uncovers more granular effects on service levels. In this subsection, we discuss two additional effects

[14] A comprehensive explanation is available in Appendix B.3.

and complement our understanding of how and why the introduction of AVs may affect service levels.

Distance between zones. When a high-demand zone and a low-demand zone are close, their service quality will interact. The low-demand zone can benefit from the supply in the high-demand zone, while the high-demand zone will be harmed due to the loss of supply for serving requests in the low-demand zone. For instance, as shown in Figure 3.6a, the Upper East Side (north) in Upper Manhattan is a high-demand zone with a demand of 278 trips per km^2, while Alphabet City in Midtown Manhattan has a relatively lower demand with a density of only about 113 hourly trips per km^2. However, the service level is reduced by 18.9% in Alphabet City, whereas the service level degradation is 24.4% in the Upper East Side (north), as shown in Figure 3.1. Because Alphabet City is close to more high-demand zones in Midtown Manhattan, the requests originating from Alphabet City can be timely fulfilled by the supply in its nearby zones. In contrast, the Upper East Side (north) is far from the most high-demand zones in Midtown and Downtown Manhattan and closer to the low-demand zones in the north. Vehicles in the Upper East Side (north) may be dispatched to serve the requests in the nearby low-demand zones, reducing its service level.

Demand imbalance. Demand imbalance, defined as the density of demand arriving in a zone (i.e., incoming trips) minus the density of demand leaving that zone (i.e., outgoing trips), can also influence the spatial inequality of service level. When the incoming trips are more than the outgoing trips in a zone (e.g., business areas during the morning rush hour), this zone is oversupplied and has a higher service level because more vehicles with riders are arriving there, but fewer riders want to leave there. Similarly, when the outgoing trips are more than the incoming trips in a zone (e.g., residential areas during the morning rush hour), the zone is undersupplied, and riders can hardly find an available vehicle. To see this, we repeated the experiment with another dataset for the morning rush hour (7 AM - 9 AM) during the workdays in January 2020. The results are shown in Figure 3.9. We can see that the demand distribution during the morning is similar to that during the middle



(a) The demand density. (b) The demand imbalance. (c) The service levels when $N_A = 8000$.

Figure 3.9: To show the effect of demand imbalance on service levels, we repeated the experiment by using the dataset for the morning rush hour (7 AM - 9 AM) during the workdays in January 2020.

of the day, where Manhattan and the airports have the highest demand density. However, compared with Figure 3.6b, Figure 3.9b shows the demand is much more imbalanced during the morning because many requests are from residential to business areas. For example, as shown in Figure 3.9b, the business areas such as the financial district, the Midtown, and the north of Queens borough have more incoming than outgoing trips. As a result, for these areas, the service levels shown in Figure 3.9c are higher than the service levels shown in Figure 3.8a.

3.7 Conclusion

We model a ride-hailing mixed fleet management problem in the presence of AVs using a queueing model. Specifically, a profit-maximizing platform chooses how to serve requests and maximize its profit through its fleet of AVs and HVs, while human drivers decide to join the platform based on their equilibrium earnings. This model is simple enough to derive the platform's optimal dispatch policy yet still rich enough to illustrate the critical interaction

of dispatch policies, driver wage equilibrium, and geospatial service levels.

We derive two main insights from this model. First, we reveal that introducing AVs may lead to decreased service levels (which can be interpreted as a measure of service reliability in the ride-hailing market). We explain that when the pay ratio is not too small, the platform chooses a profit-maximizing policy that prioritizes AVs, decreasing HV revenue and driving them out of the market. More HVs leave than AVs are added, which lowers the total supply and the service level. Prioritizing AVs is optimal because serving a ride with an AV has a higher profit margin, and the corresponding added profit because of prioritization is worth the loss of demand and revenue due to the decreased total supply. Second, we show that the service level deterioration may be particularly severe at remote or low-demand locations. Indeed, it is also profit-maximizing for the platform to concentrate its AVs on high-demand areas. This leaves HVs in charge of the low-demand areas, but HVs are also leaving the market, which particularly negatively affects the low-demand area service level.

These findings are confirmed in a highly detailed large-scale simulation in New York City, which corresponds to a much more realistic setting. We also show that our results are robust to a variety of settings and passenger preferences. Intuitively, the loss of service level and its spatial imbalance are a first-order consequence of three factors: (a) the platform maximizes profit, (c) AVs have a higher profit margin per ride than HVs, and (c) the platform is able to prioritize the use of AVs (through direct control or incentives). This is why we expect our findings to be robust to a variety of settings that we have not necessarily explored and could potentially even occur if the AVs and HVs are operated by different platforms, as the AV platform should still concentrate their vehicles in higher demand areas and use its profit advantage to cannibalize demand from the HV platform and push human drivers to lower demand areas.

We hope this work sheds on the perhaps counter-intuitive effects of the rise of AV fleets and helps foreshadow and prevent its potentially mixed impact on on-demand transportation.

CHAPTER 4

Supply Prioritization in Hybrid Marketplaces

4.1 Introduction

The growth of the gig economy and the rise of labor marketplaces have enabled many industries to use independent contractors instead of traditional employee workforces. The question of worker categorization is particularly relevant for ridesharing platforms. Recent debates demonstrate that some drivers prefer the flexibility of contractor models, whereas others would prefer to become employees (Solis, 2021). Interestingly, the two models can be combined, and companies can use hybrid workforces, a mix of employees and contractors. For instance, e-commerce companies such as Amazon use a combination of traditionally employed delivery drivers with crowdsourced alternatives, allowing them to expand their logistics network and lower costs (Dolan, 2022). A survey conducted by Harvard Business School and Boston Consulting Group reports that almost 90% of business leaders consider digital technology platforms and using a hybrid workforce as imperative tools for their competitive advantage (Fuller et al., 2020). The use of contractors to supplement existing employees has also become particularly attractive during the COVID-19 pandemic: workers value flexible schedules, and firms face increasingly erratic demand patterns (Fuller et al. (2020), Ogg (2021)).

The use of hybrid workforces is not limited to employee/contractor models, however. For example, consider the case of the deployment of autonomous vehicles (AVs). Advantages of AVs include a substantial decrease in operating cost per mile (Hazan et al. (2016), Fagnant

and Kockelman (2018), Litman (2023)) and improved reliability stemming from complete control of the vehicle. The latent potential of this new technology has led several ride-hailing platforms such as Lyft and Uber (Uber (2016), Lyft (2024)), but also others such as Google and Amazon, to invest in self-driving cars (Dave and Jin (2021), Hawkins (2021)). Some ride-hailing platforms, such as Lyft (Lyft, 2021), anticipate that they will need to operate a mixed fleet, combining autonomous vehicles with a human driver contractor marketplace.

Regardless of the specific context, there are commonalities that are fundamental to hybrid marketplaces. The acquisition and operation of autonomous vehicles is not unlike hiring employees. A firm commits and economically sustains a staffing level of *private agents* (e.g., the number of autonomous vehicles or employees) regardless of the actual revenue generated by these workers. A firm can also source its supply from *flexible agents* or contractors who are self-scheduled, can freely join the platform, and are paid based on the amount of work or service they deliver. In the case of platforms, typically, the firm takes a commission and must ensure that these workers are incentivized to join the market. While flexible and private supply agents have intrinsic differences, their presence is a general feature of hybrid marketplaces. This is true for a wide range of applications, including ride-hailing, freelance labor marketplaces, food delivery services, or short-term home rental services like AirBnB.

While these two types of supply have different economic structures that may make the operations of a hybrid workforce advantageous (Lobel et al., 2024), a hybrid workforce is not without its challenges. A firm can set its level of private supply precisely, but this represents a significant commitment as hiring or purchasing decisions have high fixed costs and lead time. The firm does not directly set the flexible supply, but it must incentivize flexible supply agents to join the marketplace through proper incentives. The hourly cost is another source of distinction. In some settings, the hourly costs of private supply can be higher than the typical hourly flexible supply earnings. For example, Uber estimates that classifying its workers as employees would increase costs significantly (Khosrowshahi, 2020). In other situations, private supply can be less expensive. For example, the capital

and variable hourly costs of autonomous vehicles are expected to be less expensive than the typical hourly earnings of human drivers (Fagnant and Kockelman (2018), Litman (2023)). The firm thus faces a complex management problem characterized by the interplay between different economic incentives and cost structures: Giving priority to the flexible supply can incentivize a desired staffing level, but this might be too costly. Prioritizing the private supply might lower the firm's costs, but it can reduce the overall supply. Not acknowledging this interplay and ignoring the fundamental differences between the different kinds of supply agents can lead to marketplaces that are chronically over- or under-supplied, thus hurting the firm's profitability.

In this work, we consider a profit-maximizing firm that has access to both private and flexible supply. The firm considers the advantages and limitations of its supply alternatives to make two crucial decisions: staffing and supply management. The firm must decide the supply mix (staffing): *should it operate both types of supply, or only one?* How many private supply agents should the firm hire? Given the presence of both private and flexible supply, the firm must make decisions (e.g., pricing and matching) that affect the revenue generated by each supply type (supply management): *should the firm treat them differently or equally?* In the context of ride-hailing, should the firm use pricing and dynamic matching policies that treat employees and contractors differently? For example, given that employees are a "sunk cost," the firm may want to prioritize them when possible instead of paying for contractor work. However, this choice is not evident as prioritization typically introduces inefficiencies in the marketplace. In ride-hailing, it may lead to higher wait times for the customers, as a policy that prioritizes employees could favor dispatching an employee even if they are not the closest available driver. Additionally, de-prioritizing contractors may lower the number of contractors willing to join the marketplace. Therefore, the management of hybrid workforces may necessitate a change in the traditional management of marketplaces and companies.

We aim to answer these questions and elucidate the complex interactions between marketplace staffing and supply management policies. While hybrid marketplaces present unique

challenges, we will show that there is a general, tractable structure to this problem. Based on the characteristics of the firm, we will be able to quantify the advantage of managing a hybrid workforce, and in particular to show the key role of supply prioritization in hybrid marketplaces.

In our model, the firm first makes its staffing decision, i.e., determines the amount of private supply it needs and whether it should also use flexible supply. It also sets its policies: all the rules, processes, and algorithms that link available supply to expected revenue, such as pricing and online matching policies in a ride-hailing application. These policies may consider the type of supply available and even prioritize one supply type in some way, such as giving private agents more work. The firm also sets its commission rate—the fraction of revenue it will take from its flexible agents. Based on these first-stage decisions, the flexible agents join the marketplace by gauging their earnings from joining (based on the pay ratio) against an outside option. Therefore the firm can optimize the level of private supply, but flexible supply is set in equilibrium.

4.1.1 Main Contributions

A general axiomatic approach. We develop a general modeling framework that captures a wide range of applications and firms' decisions. A main challenge is that every market has its own set of unique characteristics and policies available to firms. This can render the study of supply prioritization policies highly dependent on a given application. For example, such policies would correspond to the choice of pricing and dispatch algorithms in ridesharing or the way properties are listed and presented to customers in AirBnB. Modeling a firm's operations explicitly (for example, with a queueing model) is typical in the literature but it makes results and insights application-dependent. Instead, we focus on macro-level quantities implied by the firm's operations and implemented policies. We study how the firm staffs and prioritizes its supply agents, not by analyzing specific policies but by assessing the revenue achieved by the different supply types. We introduce various realistic axioms on

the behavior of the achievable revenues, and prove general results that lead to managerial implications that seem to be fundamental for hybrid marketplaces. A critical insight is that complex supply management policies can be modeled generally by focusing on the outcome they induce on the macro-level quantities rather than the details of their implementation.

Optimal staffing decisions. One of our main contributions is to characterize the optimal staffing decisions of the firm. The firm needs to choose its most profitable option: being a traditional flexible-only marketplace, investing in private supply and operating a hybrid marketplace, or using private supply only. The most simple case is when private supply is cheap, specifically when the hourly cost of private supply is lower than the flexible supply reserve wage (their best outside option). In that case, it is optimal for the firm to operate private supply only. If private supply is expensive, the optimal staffing policy depends on how precisely the firm can control its flexible supply marketplace. The firm should stay flexible-only if it can precisely set its commission rate so that the flexible supply level is optimal for the market—neither over-supplied nor under-supplied. These two first cases are reassuring as they correspond to the most common staffing models (hybrid staffing is still rare in practice). However, most marketplaces rarely adapt their commission rate to current supply conditions. Uber and Lyft are chronically over-supplied or under-supplied as the flexible supply best outside option can change over time. The platforms cannot adapt their commission fast enough to compensate. In that case, we show that hybrid marketplaces may be optimal, even if private supply is expensive. We show that three factors contribute to making hybrid marketplaces particularly advantageous: (a) a moderate cost of private supply, (b) the imbalance of the flexible supply market (e.g., significantly over-supplied or under-supplied), and (c) the access to complex supply management policies that can significantly prioritize supply. As discussed below, our central insight is that this last condition—access to supply prioritization—is crucial for the viability of a hybrid marketplace.

The key role of prioritization policies. Being able to prioritize one supply type is essential for hybrid supply to be optimal. We present a complete characterization of "equal-

treatment" policies, e.g., policies that do not prioritize one type of supply over the other. These results show that a hybrid supply staffing strategy cannot be optimal without prioritization. Nonetheless, a firm using prioritization strategies has to balance complex effects. Suppose that a firm chooses to invest in expensive private agents. It can then try to prioritize them to increase their productivity and make this investment worthwhile. However, this would come at the expense of flexible supply, reducing their earnings. Flexible supply would then exit the market, impacting the potential revenue that the firm can achieve and potentially erasing the positive effect of prioritizing private supply. A surprising reverse strategy can also work: the firm could decide to de-prioritize the expensive private supply to increase the revenue of flexible supply. This approach may seem questionable as private supply would have low productivity, but the added flexible supply revenue could grow the amount of flexible supply available in equilibrium. Actually, if demand is elastic enough to the availability of supply, this may trigger a positive "snowball effect". The extra flexible supply revenue from prioritization leads to an increased flexible supply, leading to a further increase in demand/revenue, leading to a further increase in flexible supply. Therefore, at equilibrium, a small amount of flexible supply prioritization can potentially add a lot of flexible supply and revenue to the market. This effect can be strong enough to compensate for the low productivity of the private supply and therefore increase profit overall. We confirm this intuition with a sharp result. In a hybrid marketplace, if the flexible supply market is "over-supplied" (even slightly), then it is always optimal to use strategies that prioritize private supply. Conversely, if the flexible supply market is "under-supplied", it is optimal to prioritize flexible supply.

The limitations of prioritization. Even if prioritization strategies are optimal when both private and flexible supply is present, it is not clear whether it is worth it for the firm to pay for the expensive private supply. For hybrid staffing to be optimal, the increase in profit from prioritization needs to be high enough to compensate for the private supply costs. Two limitations of prioritization strategies can prevent this from being the case. First, we

show that the only way prioritization policies can increase profit significantly is if the flexible supply market is not too close to being perfectly balanced. That is, the flexible market needs to be over-supplied or under-supplied *enough* for hybrid supply to be optimal. Second, using prioritization policies also introduces inefficiencies in the marketplace. Consider the example of the prioritization of autonomous vehicles in ride sharing. A ride request may have a human driver nearby (flexible supply). However, the platform may instead choose to assign a further away autonomous vehicle if it is trying to prioritize these vehicles. This increase in wait time would increase the probability of customer cancellation and, therefore, reduce the expected revenue that the firm would obtain from this match. We confirm this intuition by showing that prioritization policies may introduce inefficiencies that can reduce the expected revenue of the firm. Our final theorem quantifies these effects and establishes the conditions for the optimality of a hybrid marketplace: the firm needs to be able to prioritize supply enough, with a low enough level of prioritization inefficiency, and a flexible supply market that is imbalanced enough.

The remainder of this chapter is organized as follows. In Section 4.1.2, we review the related literature. In Section 4.2, we present our axiomatic framework with minimal assumptions for a hybrid marketplace. In Section 4.3, the equal treatment policies are characterized and utilized as a foundation to analyze the more complex prioritization policies. We discuss the optimality of prioritization given constant private supply in Section 4.4. And in Section 4.5, we show that a hybrid marketplace may be optimal only with a prioritization policy. We conclude in Section 4.6.

4.1.2 Related Literature

Blended workforce literature. Our work is related to the emerging literature on blended workforce in the gig economy. Dong and Ibrahim (2020) study a cost-minimizing staffing problem where the manager has to decide how many full-time employees and flexible contractors to staff in order to balance operating costs, varying demand patterns, and supply

uncertainty. They derive staffing policies based on fluid and stochastic formulations within a queueing framework. Lobel et al. (2024) investigate how the firm should staff its operations with employees and contractors with unknown demand distribution. The authors consider two scenarios where the company may or may not adjust the contractor wage after observing the state of the world. They show that contractors' flexibility can help the firm choose the optimal contractor utilization despite demand uncertainty. In this sense, using employees can perform arbitrarily worse than using contractors. He and Goh (2022) use a model of last-mile parcel delivery to elucidate how the demand should be allocated between employees and freelancers. They illustrate that the influence of acquiring freelancers on the system's profit depends on the mean and variance of the cross-network effect, which depicts how demand increases with the size of the freelancers pool. Chakravarty (2021) studies a ride-hailing platform with blended driver capacity and compares two particular forms of demand rationing: preferential rationing and driver-agnostic rationing. In addition, some other studies also examine how to minimize production costs with mixed labor force available (e.g., Kesavan et al. (2014) and Bhandari et al. (2008)).

Additionally, Hu et al. (2023) develop a queueing model to investigate the question of whether long-term workers in a gig economy can benefit from being reclassified. The authors point out that undercutting and overjoining are two fundamental issues when all the workers have the same classification. To offset these issues, they propose a hybrid mode with a discriminatory scheme where long term workers are prioritized over ad hoc workers. In contrast, we focus on supply prioritization itself and study optimal prioritization strategies. Krishnan et al. (2022) is the only paper that, to our knowledge, studies an actual implementation of a blended workforce using prioritization. This work describes how the ridesharing firm Lyft separates the drivers into "priority drivers," prioritized by the matching system to have higher earnings, and regular "flexible drivers." While the priority drivers are not precisely "private" agents, this paper shares similarities with our work. It gives a compelling practical example of the tradeoffs of prioritization, with its

revenue inefficiencies and interaction with flexible supply equilibrium. Our work generally contributes to this burgeoning literature by proposing a more general model which allows for any form of prioritization and unspecified operational decisions. We are the first to describe the joint optimal staffing and operational policy decisions in this broad context. Our general setting enables us to achieve a broad understanding of the interactions between the firm's supply management policies and the state of its flexible supply market.

Self-scheduling capacity literature. More broadly, our work relates to the literature on self-scheduling capacity, as our flexible supply agents are self-scheduled and join the firm based on their equilibrium earnings. Gurvich et al. (2019) study the capacity management of a ride-hailing market with self-scheduling drivers. They illustrate that the firm may incur extra costs and the customers may receive less service as a result of self-scheduling. Cachon et al. (2017) discuss different pricing schemes with self-scheduling providers and uncertain demand. They show that surge pricing can benefit all the stakeholders with self-scheduling capacity. Cachon et al. (2021) investigates how online service platforms should choose between the centralized and decentralized control of price. The authors find that under a simple commission structure, this decision relies on the competition among the servers and how much servers value setting prices independently. Ibrahim (2018) considers a staffing queueing problem with a random number of servers and impatient customers. It proposes making delay announcements to control customer abandonment behavior and mitigate the cost due to the uncertainty of servers. In the paper of Afèche et al. (2023), the authors study a ride-sharing problem of matching riders with self-scheduling drivers by using a game-theoretic fluid model. They point out that despite excess supply, it may be optimal for the company to reject requests at a low-demand location to induce repositioning to a high-demand location. Taylor (2018) adopts a queueing system to analyze how the delay sensitivity and agent independence affect the optimal price and wage of the firm. The paper of Benjaafar et al. (2022) examines how the labor welfare is influenced when firms expand the labor pool and impose a wage floor. And Hu and Zhou (2020) discusses the performance

of a fixed commission contract with respect to demand and supply elasticity. Our study uses the concept of self-scheduling capacity through the notion of flexible supply equilibrium. However, we also incorporate private agents, and focus on the interaction between them and the resulting prioritization implications. a

4.2 Model

We consider a firm that uses two types of supply to generate revenue. *Private supply* is fully controlled by the firm, while *flexible supply* must be secured through proper incentives. Our goal is to understand how a profit-maximizing firm should choose its supply mix and how it should use these two types of supply to generate revenue. For example, the firm could use supply management policies that prioritize the use of its private or flexible supply. In order to capture a wide variety of realistic settings, we study these questions in the context of an axiomatic framework with minimal assumptions.

We consider the firm's operations over a relatively long time-frame (e.g., a year). Nonetheless, this duration is short enough for the labor market conditions to be considered stationary. Specifically, the cost of private supply and the equilibrium earnings of flexible supply will be assumed stationary, as discussed later. At the beginning of the time-frame, the firm chooses how to operate its supply. The firm can choose any feasible operational policy that it can implement and influence how its supply generates revenue. For example, a ride-hailing platform can choose the ride pricing policy, the dynamic matching algorithm that matches drivers with requests, or even its routing suggestions to the available drivers. Additionally, the firm decides how much to invest in its private supply (e.g., how many employees to hire or autonomous vehicles to invest in). The firm wants to find the decisions that will maximize its expected earnings for the time-frame, considering that the customers and the flexible supply will react to its decisions. To model this choice, we will focus on macro-level quantities such as the expected profit/revenue of the firm or the number of supply-hours

over the time-frame.

Supply. Let N_F and N_P be the total expected number of available hours of flexible and private supply over the time horizon, including both the idle hours and the hours when they serve customers. In particular, N_F represents the available hours during which flexible supply agents are willing to work for the firm — as we explain below, N_F is endogenous and decided by flexible agents in equilibrium. The firm can control N_F indirectly via incentives such as its compensation policy. N_P represents the available hours of the private supply. N_P is either optimally selected by the firm at the beginning of the time period, or may be exogenously given (for example, a ride-hailing company may already own a fleet of autonomous vehicles). In particular, the expected supply-hours N_P and N_F may be arbitrarily distributed over the time horizon based on the scheduling decisions of the firm, and the self-scheduling choices of flexible supply.

Policies and revenue. The firm manages revenue and supply using various operational decisions and algorithms, such as matching, pricing, and routing in the context of ride-hailing. We will refer to these choices as the *policy* of the firm. Let $\pi \in \Pi$ be the policy chosen by the firm, where Π is the set of all policies that the firm can choose from to use over the time horizon. We note that Π is general, and may contain any complex policy that this specific firm can use within its technological limits. Over the time horizon, the firm will generate an expected revenue, which we assume can be attributed to its two types of supply: R_P, R_F are the expected revenue earned because of private and flexible supply, respectively, and $R_P + R_F$ is the total revenue of the firm. For example, in ride-hailing, these revenues would correspond to the expected sum of the prices customers are paying for being served by either supply source. This revenue varies with the choice of policy π and the available supply; we use the notation $R_P^\pi(N_P, N_F)$ and $R_F^\pi(N_P, N_F)$ to highlight this dependence. Intuitively, given fixed (N_P, N_F), a pricing strategy may increase or decrease both R_P, R_F depending on the demand response. We do not explicitly model the variety of specific operational policies or the demand response to these policies. Instead, we encode

their implied outcomes in the revenue functions, $R_P^\pi(N_P, N_F)$ and $R_F^\pi(N_P, N_F)$. Note that R_P and R_F are typically not "independent": for example, if a ride-hailing firm chooses a matching strategy that prioritizes private agents at the expense of flexible agents, this would increase R_P and decrease R_F.

These functions must satisfy some natural properties. First, if one supply type is unavailable, there should be no corresponding revenue. Second, the total revenue the firm can garner is upper bounded. Indeed, the maximum willingness to pay for service in a given time horizon is trivially upper-bounded, no matter how perfect the service is or how large is the available supply. Formally, there exists $M > 0$ such that for any policy $\pi \in \Pi$ and $N_P, N_F \geqslant 0$ the firm's revenue functions verify:

$$R_P^\pi(0, N_F) = 0 \quad \text{and} \quad R_F^\pi(N_P, 0) = 0; \quad \text{and} \quad R_P^\pi(N_P, N_F) + R_F^\pi(N_P, N_F) \leqslant M.$$

Flexible supply equilibrium. As with many service firms (e.g., ride-hailing or food delivery), we assume that the revenue from the service provided by a flexible agent is shared between the firm and the agent, with a pay ratio $\gamma \geqslant 0$ (e.g., $1 - \gamma$ is the commission rate of the firm). The flexible supply agents' total expected earnings over the time horizon are γR_F, and the firm's share is $(1 - \gamma)R_F$. Note that we allow $\gamma > 1$, which would correspond to subsidizing this supply source. We can now define the average hourly earnings of flexible supply as:

$$\begin{cases} \gamma \cdot \frac{R_F^\pi(N_P, N_F)}{N_F} & \text{if } N_F > 0; \\ 0 & \text{if } N_F = 0. \end{cases}$$

Following the empirical work of Hall et al. (2021), we assume that the flexible supply market is very elastic, that flexible supply decides to enter and exit the market based on its average earnings. Formally, let $r > 0$ be the reserve earnings of flexible agents for the time horizon, i.e., what they could make per hour not working for the firm. Suppose that $N_F > 0$ then flexible supply average hourly earning within the system is $\gamma R_F / N_F$ (where R_F is a notation shortcut for $R_F^\pi(N_P, N_F)$) If $\gamma R_F / N_F < r$, some flexible agents working for the firm are

not making enough. These agents would spend less time working for the firm or even not work at all and, therefore, N_F would decrease until $r \leqslant \gamma R_F/N_F$, or $N_F = 0$. Conversely, if $r < \gamma R_F/N_F$, more flexible agents would be willing to work for the firm, and $\gamma R_F/N_F$ would decrease until $r \geqslant \gamma R_F/N_F$. Therefore, in equilibrium, N_F must be verify:

$$\gamma R_F^\pi(N_P, N_F) = rN_F. \tag{4.1}$$

Any N_F satisfying the equation above is an equilibrium.[1] Given N_P, γ and π, we use $\mathcal{E}^\pi(N_P, \gamma)$ to denote the set of possible equilibria i.e., solutions to eq. (4.1). Note that Hall et al. (2021) finds that flexible supply takes several weeks to reach an equilibrium in ride-hailing. As we are assuming that flexible supply is in equilibrium, we need our time horizon to be large enough for flexible supply to react to the firm's decisions, yet small enough for r to be considered constant.

Profit Maximization. The firm wants to maximize its profit over the time horizon. As mentioned above, the firm keeps $(1-\gamma)R_F$ from the total revenue generated by the flexible supply. In the meantime, the firm keeps the totality of its private supply revenue; however, we assume that each private supply hour costs $C_P > 0$ to the firm. This cost may represent fixed or amortized capital costs (e.g., if private supply represents autonomous vehicles in ride-hailing) and variable costs associated with this source of supply (e.g., employees' hourly pay if private supply represents employee hours).

Given the above, the firm first chooses N_P, γ and π, and then flexible supply reacts in equilibrium by setting $N_F \in \mathcal{E}^\pi(N_P, \gamma)$. The firm anticipates the flexible supply equilibrium and chooses a policy that maximizes its profit:

$$\max_{\pi \in \Pi, \gamma \geqslant 0, N_P \geqslant 0} R_P^\pi(N_P, N_F) + (1-\gamma)R_F^\pi(N_P, N_F) - C_P N_P, \quad \text{s.t} \quad N_F \in \mathcal{E}^\pi(N_P, \gamma). \tag{4.2}$$

In Problem (4.2) the firm can completely control all the levers at its disposal, i.e., the policy, the flexible supply hours, and the pay ratio. However, it is not evident that the firm can

[1] Formally, an equilibrium satisfies: $N_F > 0 \implies \gamma R_F^\pi(N_P, N_F) = rN_F$. In our setting, this is equivalent to eq. (4.1).

perfectly set N_P and γ, depending on the setting of interest. This work will study (4.2) when N_P and γ are either optimized or exogenous, for the reasons described below.

Choice of N_P Recall that N_P is the available hours of private supply agents. Assuming that the firm can optimize N_P is a good model if the time horizon is long enough for the firm to change its staffing strategy entirely. As discussed previously, the horizon is typically long if C_p and r are relatively stationary compared to the time for the firm to adapt its private supply staffing. Nonetheless, there are settings where it is slow to change private supply or when C_p and r change quickly. For example, hiring and firing employees can be slow, and in our example of autonomous vehicles in ride-hailing, firms may build an autonomous vehicle fleet slowly over time (Litman (2023)). Additionally, C_p and r have also changed quickly during the Covid-19 pandemic, and ride-hailing platforms suddenly experienced a severe lack of supply. In these settings with a shorter time horizon or fixed private supply, considering that N_p is fixed and exogenous is a preferable model: the firm enters the time period with a certain number of private supply hours available. The question is how to use it best to maximize profit. Consequently, we will study both the fixed and optimal N_P cases, as we believe they offer complementary insights.

Choice of γ We have a similar modeling issue with the choice of the flexible supply pay ratio γ, or equivalently the firm's commission rate $1-\gamma$. Indeed, it can often be challenging for a marketplace to update its pay ratio, especially in a competitive landscape. For example, in the ride-hailing industry, the pay ratio is usually set between 75% and 85%[2]. This pay ratio did not change during the Covid-19 pandemic despite essential changes in the supply market. Therefore, it may be preferable to consider that γ is fixed and exogenous. Additionally, we will show in the study that the ability to set an optimal γ is equivalent to the ability to set the optimal flexible supply amount in the market. However, marketplaces often go through long-term over-supplied or under-supplied conditions, suggesting that modeling γ as optimal may not be a realistic model. Nonetheless, the case where γ is set optimally by the firm is

[2] `https://ride.guru/content/resources/driver-payout-take-home`.

also crucial to understanding longer-term horizons. We will, therefore, study both the fixed and optimal γ cases, as we believe they also offer complementary insights.

4.2.1 Problem Reformulation Via the Achievable Revenues Set

The profit maximization Problem (4.2) is not particularly easy to manipulate as the set of policies Π is a rather abstract object. Instead, we will prove that a more manageable formulation that only involves interpretable quantities is equivalent. This reformulation optimizes over the space achievable revenue outcomes of the policies rather than the space of policies themselves.

For all supply hours pairs N_P, N_F , we define the *achievable revenues set*:

$$\mathcal{AR}(N_P, N_F) \triangleq \{(R_P, R_F) \in \mathbb{R}^2 : \exists \pi \in \Pi,\ R_P = R_P^\pi(N_P, N_F),\ \ R_F = R_F^\pi(N_P, N_F)\}.$$

That is, $\mathcal{AR}(N_P, N_F)$ is the set of all the revenue pairs that are achievable by policies in Π, given that the available supply hours are N_P and N_F. With some abuse of terminology, we will refer to the elements of the achievable revenues set as policies. For fixed $\gamma \geqslant 0$, and $N_P \geqslant 0$, remember that the firm profit maximization problem (4.2) is:

$$\max_{\pi \in \Pi} R_P^\pi(N_P, N_F) + (1-\gamma) R_F^\pi(N_P, N_F) - C_P N_P, \quad \text{s.t} \quad N_F \in \mathcal{E}^\pi(N_P, \gamma). \tag{2'}$$

We can now recast Problem (4.2) in terms of R_P, R_F and N_F.

Lemma 3 (Problem Reformulation)**.** Problem (2') is equivalent to the following optimization problem:

$$\begin{aligned} \max_{R_P, R_F, N_F} \quad & R_P + (1-\gamma) R_F - C_P N_P \\ \text{s.t.} \quad & r N_F = \gamma R_F, \\ & (R_P, R_F) \in \mathcal{AR}(N_P, N_F). \end{aligned} \tag{4.3}$$

All the proofs of the study are available in the appendix. The lemma establishes that any optimal policy in Problem (2') corresponds to an optimal set of supply hours and revenues

R_P, R_F, N_F in Problem (4.3). Conversely, given any optimal solution of Problem (4.3), we can reconstruct another solution that is feasible and optimal in Problem (2'). An immediate corollary of Lemma 3 is that the reformulation also works if the firm optimizes over γ and N_P. This reformulation will be the focus of the rest of the study, as it dramatically simplifies the optimization problem. It reduces the dimensionality of the problem, from the more abstract space of policies Π to $\mathbb{R}^3$; and the objective function is a simple linear combination of the three variables. However, note that the achievable revenues sets $\mathcal{AR}(N_P, N_F)$ now hide most of the complexity of the problem as they are equivalent to a general constraint on the feasible space. Nevertheless, we will impose natural structural properties on $\mathcal{AR}(N_P, N_F)$, allowing us to characterize the solutions of Problem (4.3).

A first structural property that we will impose on the achievable revenues set, and perhaps the most crucial assumption of this study, is its *symmetry*. We first present the assumption in formal terms and then describe its justification. We will assume that the two supply types are symmetric in how the firm can use them to generate revenue. Any total revenue $R_P + R_F$ that is feasible for a firm when it has some supply N_P, N_F available can also be achieved with only one type of supply or any other supply mix, but the same total number of supply hours. To formalize this, we first define $\mathcal{AR}(N)$ to be the set of achievable revenues of a "single-type" policy only using flexible supply for any $N \geqslant 0$:

$$\forall N \geq 0,\, \mathcal{AR}(N) \triangleq \{R_F \,|\, (0, R_F) \in \mathcal{AR}(0, N)\}.$$

Then the symmetry condition can be stated as follows.

Assumption 1 (Symmetry of supply types)**.** The feasible total revenues that the firm can achieve is a function of the total supply hours available. It is independent of supply mix.

$$\forall N_P, N_F \geqslant 0,\ \{R_P + R_F \,|\, (R_P, R_F) \in \mathcal{AR}(N_P, N_F)\} = \mathcal{AR}(N_P + N_F).$$

This leads to a much more intuitive definition of $\mathcal{AR}(N)$[3]: it is the set of achievable total revenue with N total supply hours, regardless of the supply mix.

[3]Note that $\mathcal{AR}(N)$ and $\mathcal{AR}(N_P, N_F)$ are different.

This assumption is particularly strong; we discuss next that it holds when there is no fundamental operational difference between the two types of supply, except for their staffing mechanism: they are "symmetric". A simple way of understanding what we mean by symmetry is to compare two situations. *(i)* One in which the firm has access to N_P and N_F supply hours from each type, and *(ii)* another with the same total supply hours but only flexible agents, that is, $N'_P = 0$ and $N'_F = N_P + N_F$. Consider a given policy in situation *(i)* with corresponding average revenues R_P and R_F, and total revenue given by $R_P + R_F$. If the two types of supply are operationally equivalent (e.g., customers are indifferent to the supply type, the supply agents work with the same efficiency, etc.), then the firm can actually replicate this policy in situation *(ii)* with only flexible supply. Indeed, even if only flexible supply is available, the firm can randomly pretend that some flexible agents are instead private and use the previous policy. Because the agents are operationally equivalent, this new policy will achieve the same total revenue $R'_F = R_P + R_F$. Similarly, given a policy in situation *(ii)*, the firm can replicate it in the mixed-supply world by simply ignoring the type of supply and applying the same policy. Therefore, if we suppose that the supply types are "symmetric", Assumption 1 must be true, and this is the intuition behind the assumption.

This symmetry assumption is often not exactly true in practice. In ride-hailing, riders may be reluctant to use autonomous vehicles, and autonomous vehicles may only be able to operate in some areas of a city. In that case, the two types of supply are not symmetric, and the firm would not be able to achieve the same revenue with autonomous vehicles as it achieves with human drivers. Or on the contrary, a firm using employees may have more control over their work hours than a firm using self-scheduled contractors. This increased control could make employees more efficient and generate more revenues than contractors. However, all of these asymmetric effects are application specific. Using the strong Assumption 1, we can isolate the effect of the staffing and control policies of the supply type from the inherent application-dependent superiority of one type of supply. Assuming that agents

are interchangeable, we exclude such external factors to focus on the main intuition behind the optimal staffing and control of private and flexible supply. Nonetheless, it is easy to model specific situations with supply asymmetry using the tools of this study, as we show in Appendix C.6 of the Appendix where we study a special case and show that most of the results and insights of the study still hold in that case.

4.2.2 Definition of Equal Treatment and Prioritization Policies

Now that our model and main assumption are defined, we are going to partition the firm policies into three categories: private supply prioritization, flexible supply prioritization and equal treatment. Because our reformulation focuses on achievable outcomes, we will not directly describe what it means for a policy to prioritize a supply type. Instead, we will define prioritization based on a policy's impact on the two supply-type expected hourly revenue. The firm prioritizes private supply if the chosen policy leads to a higher expected revenue per supply hour R_P/N_P than flexible supply, and flexible supply prioritization happens if R_F/N_F is higher. There are many realistic settings where a firm can easily prioritize one type of supply: for example, a ride-sharing matching policy can easily match to private agents first, and Krishnan et al. (2022) describe a practical implementation. Naturally, prioritization has a counterpart: *equal treatment*. A policy that satisfies equal treatment leads to the same average hourly revenue for both types of supply (when the flexible supply is in the market). This definition works well when the two types of supply are present ($N_F, N_P > 0$), but the special case $N_F = 0$ needs a little more care, as described in our formal definition of prioritization:

Definition 2 (Prioritization and Equal treatment)**.** For $N_P > 0$ we say that $(R_P, R_F) \in \mathcal{AR}(N_P, N_F)$

1. (Prioritizing Flexible Supply.) prioritizes flexible supply if and only if

$$N_F > 0 \text{ and } \frac{R_P}{N_P} < \frac{R_F}{N_F},$$

2. (Prioritizing Private Supply.) and that it prioritizes private supply if and only if

$$N_F > 0 \text{ and } \frac{R_P}{N_P} > \frac{R_F}{N_F}, \quad \text{or} \quad N_F = 0 \text{ and } \frac{R_P}{N_P} > \frac{r}{\gamma}.$$

In any other case, we say that (R_P, R_F) satisfies equal treatment and we use $\mathcal{ET}(N_P, N_F) \subseteq \mathcal{AR}(N_P, N_F)$ to denote the set of equal treatment revenue pairs.

The only potentially surprising aspect of this definition is the case $N_F = 0$. One may think that when only private supply is available, talking about prioritization is not well defined and the policy can only be equal-treatment. Nonetheless, our definition characterizes situations with $N_F = 0$ and and $\frac{R_P}{N_P} > \frac{r}{\gamma}$ as private supply prioritization. In this situation, the average revenue of the market is strictly above $\frac{r}{\gamma}$, which happens to be our flexible supply equilibrium hourly revenue (see the wage equilibrium equation (4.1)). Therefore, in an open flexible supply marketplace, the flexible supply would like to join the marketplace. The only way the platform can prevent this flexible supply from entering is to actively block it, by prioritizing private supply and dropping the potential flexible supply revenue under the wage equilibrium or just shutting down the flexible marketplace entirely. These types of "blocking" policies are evidently not equal-treatment, so we classify this case as private supply prioritization. This choice will also significantly simplify our main results.

Equal treatment will play a key role in our study of prioritization. These policies are a natural benchmark; all policies that do not consider the supply type are equal treatment policies. Section 4.3 study these policies in detail and describe what is achievable by the firm without prioritization. For example, in a ride-hailing market, equal treatment can emerge when pricing, matching and routing decisions do not depend on the agent type. And while Sections 4.4 and 4.5 focus on prioritization policies, equal treatment policies are still used as a benchmark.

The introduction of equal treatment policies leads to a natural extension of Assumption 1. Remember that we said that this assumption holds if the two types of supply are "symmetric", and the assumption states that the achievable total revenues only depend on

the total supply, regardless of the type. Suppose that we have the supply N_P, N_F, and let $R \in \mathcal{AR}(N_P + N_F)$ be an achievable revenue given the available supply. From Assumption 1, we know that this revenue R could also be achieved if all our supply was flexible, that is, with $N'_F = N_P + N_F$ and $N'_P = 0$. As a thought experiment, we consider the policy used in that case when we only have flexible supply, and we use it in the original two types setting with N_P, N_F. Because the policy does not distinguish between supply types, we expect that the expected revenue from each type is proportional to its availability, that is, $R_P = R \cdot \frac{N_P}{N_P+N_F}$ and $R_P = R \cdot \frac{N_F}{N_P+N_F}$. This is equivalent to $R_P N_F = R_F N_P$, that is, we have equal treatment as introduced in Definition 2. Therefore, we just presented evidence that any achievable revenue is also achievable in a two-type setting with an equal treatment policy, as we can use one-type policies and they become equal-treatment when used in a two-types setting. We formalize this intuition as a last structural assumption.

Assumption 2 (Equal-treatment policies can achieve any feasible revenue)**.** Given any supply $N_P \geqslant 0, N_F \geqslant 0$, any achievable revenue $R \in \mathcal{AR}(N_P + N_F)$ is achievable by an equal treatment policy.

$$\exists\, (R_P, R_F) \in \mathcal{ET}(N_P, N_F),\ R_P + R_F = R.$$

4.3 Hybrid Marketplaces without Prioritization

Now that our model is defined, we will first restrict the firm to the use of equal treatment policies. Equal-treatment policies are the most intuitive, and we will be able to derive the optimal staffing and policy decisions in this case. This analysis will serve as a foundation to study the more complex prioritization policies. For any γ and N_P, an optimal equal treatment solution must therefore solve the following optimization problem, which is the equivalent of (4.3) replacing $\mathcal{AR}$ with the more restricted $\mathcal{ET}$:

Definition 3 (Optimal equal treatment policy)**.** Given fixed $N_P, \gamma \geqslant 0$, an optimal equal

treatment policy, denoted by (N_F^E, R_F^E, R_P^E), is defined as the optimal solution to the problem:

$$\begin{aligned} \max_{N_F \geqslant 0, R_P, R_F} \quad & R_P + (1-\gamma)R_F - C_P N_P \\ \text{s.t.} \quad & (R_P, R_F) \in \mathcal{ET}(N_P, N_F), \\ & \gamma R_F = r N_F. \end{aligned} \tag{4.4}$$

4.3.1 Full Characterization of Optimal Equal Treatment Policies

A full characterization of an optimal solution to Problem (4.4) includes a complete description of the optimal mixed of supply types N_P and N_F, their revenues R_P, R_F, and the optimal firm's profit. A central quantity that we will need to this end is the maximum revenue that the firm can achieve for a given number of total supply hours N. We define the *maximum revenue function* $\overline{R}(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ as:

$$\overline{R}(N) \triangleq \max_N \mathcal{AR}(N).^4$$

Recall that $\mathcal{AR}(N)$ is the set of achievable revenues when there are N supply hours available (under Assumption 1, this is independent of the supply mix). Achieving $\overline{R}(N)$ means that the firm must use a policy that gets the most revenue from the available supply. In the context of ride-hailing, this would mean the use of an optimal pricing and matching strategy. Note that Assumption 2 implies that $\overline{R}(N)$ is always achievable with an equal treatment policy. For simplicity, we will assume that $\overline{R}(\cdot)$ is continuous.[5]

It is possible that there is no feasible solution of Problem (4.4) that is able to maintain a flexible supply market (e.g., when $N_F > 0$). This can happen if N_P is too high, or if there is generally not enough revenue in the market for the flexible supply. Our first intermediate result characterizes the situations where solutions with hybrid supply exist.

---

[4]The definition of $\overline{R}(\cdot)$ implicitly assumes that the maximum in $\mathcal{AR}(N)$ always exists for any N.

[5]Many results do not require this assumption, see the proofs for more details.

Lemma 4 (Existence of equal treatment solutions with hybrid supply)**.** Suppose Assumption 1 and Assumption 2 hold. Then, Problem (4.4) has a feasible solution with $N_F > 0$ if and only if there exists $N > N_P$ such that $rN \leqslant \gamma \overline{R}(N)$.

Intuitively, under equal treatment with hybrid supply, the average hourly revenue of each supply type equals the total average hourly revenue: $R_F/N_F = R_P/N_P = (R_P + R_F)/(N_P + N_F)$. For any total supply hours N, the latter quantity is bounded above by $\overline{R}(N)/N$. In turn, if $\overline{R}(N)/N < r/\gamma$ then flexible supply will not enter the market in equilibrium. Conversely, when $\overline{R}(N)/N \geqslant r/\gamma$ for some $N > N_P$, Assumption 2 implies the existence of an equal treatment policy with $N - N_P > 0$ flexible supply hours and r/γ total average earnings.

We now define another extremely important and intuitive quantity:

$$\widetilde{N} \triangleq \max\{N \,|\, \gamma \overline{R}(N) = rN\}.$$

$\widetilde{N}$ is the maximum total supply hours that can lead to a revenue that is consistent with the flexible agent's wage equilibrium. An interpretation of $\widetilde{N}$ is that it is highest possible amount of flexible supply in a flexible supply equilibrium: in most practical case $\widetilde{N}$ would be the expected amount of flexible supply in a flexible supply marketplace under an optimal revenue policy. Note that the "only if" part in the lemma can be equivalently stated as $N_P < \widetilde{N}$, using the continuity of $\overline{R}$. Intuitively, if $\widetilde{N} > N_P$ then it is possible that flexible agents enter the market under an equal treatment policy, as their average hourly revenue would be the same as the total average hourly revenue. However, when $\widetilde{N} \leqslant N_P$ the market is saturated by private supply agents and any equal treatment policy will lead to zero flexible agents in equilibrium. With this notation, we are now able to fully characterize the solutions to Problem (4.4):

Proposition 12 (Full characterization of equal treatment policies)**.** Suppose Assumption 1 and Assumption 2 hold. For any $0 \leqslant \gamma < 1$ and $N_P > 0$ the optimal solution of Problem (4.4) can be fully described as:

1. (Supply) $N_F^E = (\widetilde{N} - N_P)^+$ and

2. (Profit)
$$\begin{cases} (1-\gamma)\overline{R}(\widetilde{N}) + (r - C_P)N_P, & N_P < \widetilde{N}; \\ \overline{R}(N_P) - C_P N_P, & N_P \geqslant \widetilde{N}. \end{cases}$$

3. (Supply revenue)
$$\begin{cases} R_P^E = rN_P/\gamma,\ R_F^E = \overline{R}(\widetilde{N}) - R_P^E & N_P < \widetilde{N}; \\ R_P^E = \overline{R}(N_P),\ R_F^E = 0, & N_P \geqslant \widetilde{N}. \end{cases}$$

Proposition 12 completely solves Problem (4.4), but is also an natural result. Suppose that we start with $N_P = 0$, we only have flexible supply, so as we have seen before, we can achieve the best possible equilibrium revenue $R_F = \overline{R}(\widetilde{N})$, with a total equilibrium supply $N_F = \widetilde{N}$. Suppose now that we add some private supply but less than the previous equilibrium flexible supply: $N_P < \widetilde{N}$. Consider the solution where we keep the same total supply, and the private supply replaces the flexible supply, e.g., $N_F = \widetilde{N} - N_P$. We still use this total supply the same way, therefore the total revenue is unchanged: $R_P + R_F = \overline{R}(\widetilde{N})$. As a consequence, the average revenue of supply in the market is still $\overline{R}(\widetilde{N})/\widetilde{N} = r/\gamma$. The policy is equal treatment, so the flexible supply's average revenue is also $R_F/N_F = r/\gamma$: this is still a feasible equilibrium, and therefore this solution is feasible! Proposition 12 also proves that this is the optimal solution for $N_P < \widetilde{N}$. Consider now the case $N_P \geqslant \widetilde{N}$. There is too much total supply, and no equal treatment policy can generate enough revenue to satisfy the equilibrium flexible supply earnings: we must have $N_F = 0$. Consequently, the firm only tries to use the private supply in the best possible way, and the revenue is $R_P = \overline{R}(N_P)$.

Note that the only hybrid optimal solutions with $N_P > 0$ and $N_F > 0$ happen when $0 < N_P < \widetilde{N}$. And we just discussed that the firm's total revenue and total supply are independent of N_P for these solutions: we always have $N_P + N_F = \widetilde{N}$ and $R_P + R_F = \overline{R}(\widetilde{N})$.

So the revenue is unchanged when we add or remove private supply, and for each unit of private supply added, we remove precisely one unit of flexible supply in equilibrium. However, the profit may change as the supply cost is not constant. The firm pays C_P for each hour of private supply and r for each hour of flexible supply in equilibrium. Combining these facts, for each added hour of private supply in hybrid solutions, the profit changes by $r - C_P$, as can be seen in the profit closed form of Problem (4.4) for $N_P < \widetilde{N}$. It is always optimal to increase N_P if $C_P \leqslant r$ and decrease N_P otherwise, but this would lead to $N_P \geqslant \widetilde{N}$ or $N_P = 0$ and the solution would not be hybrid anymore. In summary, we just proved a crucial negative result. If we can optimize over N_P, hybrid marketplaces cannot be better than flexible-only and private-only marketplaces if we only use equal treatment policies. Formally, if we modify Problem (4.4) to allow the firm to choose the optimal N_P:

$$\begin{aligned} \max_{N_P, N_F \geqslant 0, R_P, R_F} \quad & R_P + (1-\gamma) R_F - C_P N_P \\ \text{s.t.} \quad & (R_P, R_F) \in \mathcal{ET}(N_P, N_F), \\ & \gamma R_F = r N_F. \end{aligned} \tag{4.5}$$

Theorem 6 (Optimality of single-type supply under equal treatment)**.** *Suppose Assumption 1 and Assumption 2 hold. Given any $\gamma \geqslant 0$, there exists an optimal solution to Problem* (4.5) *with either $N_P = 0$ or $N_F = 0$.*

A direct consequence is that single-type policies are also optimal if the firm can choose γ.

4.3.2 Optimality of Equal Treatment Policies

We now assume that the firm can choose an optimal pay ratio γ[6], and is not restricted anymore to equal treatment policies. In this setting, for any N_P, the optimal profit of the

[6]We discussed the relevance of this setting in Section 4.2.

firm is:

$$\begin{aligned} \max_{\gamma, N_F, R_P, R_F} \quad & R_P + (1-\gamma)R_F - C_P N_P \\ \text{s.t.} \quad & (R_P, R_F) \in \mathcal{AR}(N_P, N_F), \\ & \gamma R_F = r N_F. \end{aligned} \tag{4.6}$$

We will show that it is possible to achieve the optimal profit of Problem 4.6 using equal treatment policies. Therefore, despite the fundamental incentive differences that distinguish the supply types, the firm should not use any form of prioritization. First, to build intuition and simplify Problem 4.6, we show that choosing the pay ratio γ is equivalent to choosing the equilibrium flexible supply N_F:

Lemma 5 (Flexible γ problem reformulation)**.** Suppose Assumption 1 and Assumption 2 hold. For any N_P, Problem (4.6) is equivalent to:

$$\begin{aligned} \max_{N_F, R_P, R_F} \quad & R_P + R_F - r N_F - C_P N_P \\ \text{s.t.} \quad & (R_P, R_F) \in \mathcal{AR}(N_P, N_F). \end{aligned} \tag{4.7}$$

With the optimal choice $\gamma = rN_F/R_F$ if $R_F > 0$, $\gamma = 0$ otherwise.

The ability of the firm to set any N_F makes flexible supply very similar to private supply. Notice the symmetry in (4.7): the only difference between private and flexible supply is that private supply costs C_P whereas flexible supply has an (indirect) cost r. Now, consider the case where we fix N_P and N_F. The objective function of Problem (4.7) then simply maximizes the total revenue $R_P + R_F$, e.g., $\max\{R_P + R_F \,|\, (R_P, R_F) \in \mathcal{AR}(N_P, N_F)\}$. Furthermore, Assumptions 1 and 2 state that equal treatment can achieve any feasible total revenue given N_P, N_F:

$$\{R_P + R_F \,|\, (R_P, R_F) \in \mathcal{AR}(N_P, N_F)\} = \{R_P + R_F \,|\, (R_P, R_F) \in \mathcal{ET}(N_P, N_F)\}.$$

Therefore, we can replace $\mathcal{AR}$ with $\mathcal{ET}$ in Problem (4.7): the firm does not need to use prioritization policies.e

Theorem 7 (Optimality of equal treatment)**.** *Under Assumptions 1 and 2, for any $N_P \geq 0$, if the firm can choose γ, then an equal treatment policy can achieve the optimal profit. That is, there exists an optimal solution of Problem* (4.6) *verifying $(R_P, R_F) \in \mathcal{ET}(N_P, N_F)$.*

While we derived this result, it can be counter-intuitive. The two types of supply are fundamentally different, and N_P is "fixed", so the cost of private supply is a "sunk" cost. Therefore, it costs nothing for the firm to use private supply, while it needs to pay the flexible supply for the same work. Therefore one might expect that the firm would want to prioritize private supply as it is "cheaper" to use it. Our result proves that this is not true, as long as the firm can choose the ideal γ, which is equivalent to complete control of the flexible supply, as shown in Lemma 5. The firm should, in particular, achieve the highest possible revenue with its available supply, regardless of the contractor costs; and equal treatment policies can always maximize the revenue.

Under an optimal γ, we proved that the firm can limit itself to equal treatment policies. Suppose that the firm can also choose its private staffing N_P. We can combine the previous result with Theorem 6 to immediately show that the firm does not need to operate the two types of supply. In fact, we can show a more precise result:

Corollary 2 (Hybrid is not optimal)**.** Suppose that the firm chooses the pay ratio γ, the private supply N_P and can choose any policy. The optimal profit is:

$$\begin{aligned} \max_{\gamma, R_P, R_F, N_P, N_F} \quad & R_P + (1-\gamma)R_F - C_P N_P \\ \text{s.t.} \quad & rN_F = \gamma R_F, \\ & (R_P, R_F) \in \mathcal{AR}(N_P, N_F). \end{aligned} \tag{4.8}$$

If Assumption 1 holds, then any optimal solutions verify $N_F = 0$ if $C_P < r$ and $N_P = 0$ if $C_P > r$.

The result establishes that the firm uses the "cheapest" type of supply. The reformulation (4.7) conveys this intuition, as the two supply types only differ in their hourly cost C_P and

r; this is essentially why the firm should only use the cheapest type. This section suggests that using hybrid supply and complex prioritization strategies is unnecessary. Nevertheless, as discussed in Section 4.2, it is often a more realistic model to consider that γ is fixed and exogenous rather than optimized. As we will show in the following sections, prioritization and hybrid supply will then be essential.

4.4 Optimality of Supply Prioritization

We now consider the case where the pay-out ratio γ is fixed in $(0, 1)$ and not necessarily optimal. We will establish that equal treatment policies may no longer be optimal for profit, and we will characterize how supply should be prioritized. Specifically, this section characterizes the firm's optimal policies given any private supply N_P; we will discuss the optimal choice of N_P in the following section. Prioritization policies are significantly more challenging to study because they are application-dependent. To obtain general results with minimal assumptions, we will use a sensitivity analysis to see if the firm should deviate from the best equal-treatment policy and slightly prioritize private or flexible supply to increase profit. We will then recover a general result and fully characterize when the firm should prioritize private or flexible supply. But first, we will reformulate the optimization problem (4.3) (recall that (4.3) has fixed N_P and γ) to simplify the subsequent analysis.

4.4.1 Reformulation to a one-dimensional optimization problem

To reformulate Problem (4.3) into a one-dimensional problem, we decompose it into several stages. Recall that (4.3) is a three-dimensional optimization problem over R_P, R_F, N_F. We first fix N_F and R_F and derive the optimal private supply revenue R_P, and we then obtain the optimal N_F given R_F. This approach allows us to characterize the optimal profit as a function of R_F. For the first step, given fixed and feasible N_F and R_F, notice that it is optimal for the firm to choose a policy that maximizes the private supply revenue R_P.

Specifically, we define the *maximal private supply revenue function*, $\overline{R}_P(N_F, R_F)$, and its domain, $\mathcal{D}$, by

$$\overline{R}_P(N_F, R_F) \triangleq \max\{R_P \,|\, (R_P, R_F) \in \mathcal{AR}(N_P, N_F)\}, \quad (N_F, R_F) \in \mathcal{D} \tag{4.9}$$

where

$$\mathcal{D} \triangleq \{(N_F, R_F) \,|\, \exists R_P,\ (R_P, R_F) \in \mathcal{AR}(N_P, N_F)\}\,.$$

The domain $\mathcal{D}$ captures that the revenue R_F should be feasible with the supply N_F for some policy and some N_P, but note that it does not impose the flexible supply equilibrium. Given (N_F, R_F), the firm will always chooses a policy that guarantees $R_P = \overline{R}_P(N_F, R_F)$. We can then optimize over the feasible pairs of (N_F, R_F) in $\mathcal{D}$ that satisfy the equilibrium condition. Given R_F, N_F is uniquely determined by the equilibrium equation $rN_F = \gamma R_F$. Therefore, the optimal R_P in equilibrium is $\overline{R}_P(\gamma R_F/r, R_F)$, and we can use this to obtain the optimal profit as a function of R_F:

Proposition 13 (Reformulation to a one-dimensional problem)**.** Given $N_P \geqslant 0$ and $\gamma \in (0, 1)$, optimal solutions of Problem (4.3) verify $R_P = \overline{R}_P(N_F, R_F)$ and the problem is equivalent to:

$$\max_{R_F \text{ s.t. } (\gamma R_F/r, R_F) \in \mathcal{D}} \text{Profit}(R_F), \tag{4.10}$$

where $\text{Profit}(R_F) \triangleq \overline{R}_P(\gamma R_F/r, R_F) + (1-\gamma)R_F - C_P N_P$ is the optimal profit given fixed R_F.

We now only need to find the feasible R_F that maximizes profit. We next show that this choice is particularly interpretable. For example, recall that the optimal equal treatment policy verifies $R_F = R_F^E$ (see Definition 3). Actually, we can derive that there is no better choice of policy verifying $R_F = R_F^E$ and $\text{Profit}(R_F^E)$ is the profit of the optimal equal treatment policy. Generally, we are able to characterize the type of the optimal policy as a function of R_F:

Proposition 14. Suppose Assumption 1 and Assumption 2 hold. Then, given an optimal solution R_F of Problem (4.10), any policy achieving the optimal profit $\text{Profit}(R_F)$ is:

1. a private supply prioritization policy if $R_F < R_F^E$,

2. a flexible supply prioritization policy if $R_F > R_F^E$,

3. an equal treatment policy if $R_F = R_F^E$.

To see why the proposition holds, consider an optimal solution to Problem (4.10) with strictly less flexible revenue than the optimal equal treatment solution, $R_F < R_F^E$, but with higher profit, $\text{Profit}(R_F) > \text{Profit}(R_F^E)$. Under Proposition 14, we must have $R_P > R_P^E$, otherwise both $R_P \leqslant R_P^E$ and $R_F \leqslant R_F^E$ would imply $\text{Profit}(R_F) \leqslant \text{Profit}(R_F^E)$. Therefore, $R_P/N_P > R_P^E/N_P$ (N_P is fixed). Under equal treatment, the private and flexible supply average hourly revenues equal r/γ. However, in equilibrium, any policy preserves the same flexible supply average hourly revenue, r/γ. Hence we have $R_P/N_P > R_F/N_F$, which means that the firm prioritizes its private supply.

Proposition 14 provides a simple way of assessing whether an optimal solution is a prioritization or equal treatment policy: compare the flexible supply revenue of any policy with that of the optimal equal treatment policy. Nonetheless, deriving optimal R_F means solving (4.10), which is problem-specific and arbitrarily complex. Instead, the next subsection uses the gradient of $\text{Profit}(\cdot)$ at R_F^E to present an intuitive and simpler characterization of the optimality of prioritization.

4.4.2 Optimality of Prioritization

Suppose that the firm implements the optimal equal treatment solution and contemplates whether it should deviate in order to increase profit. For example, the firm could start to slightly prioritize its private supply. Any small change in the policy would result in a small change, dR_F, in the corresponding flexible supply revenue, R_F^E, at equilibrium. The flexible supply revenue would move from R_F^E to $R_F^E + dR_F$. From Proposition 14, we know that if this deviation leads to a higher profit, $\text{profit}(R_F^E + dR_F) > \text{Profit}(R_F^E)$, then an equal treatment policy is not optimal. In general, if $\text{Profit}(\cdot)$ is differentiable and its gradient is not 0 at R_F^E

then a prioritization policy is optimal. To this end, we make the following minimal regularity assumptions which will enable us to perform sensitivity analysis around the optimal equal treatment solution.

Assumption 3. $N_F^E, R_F^E, \overline{R}_P(\cdot,\cdot)$ and $\overline{R}(\cdot)$ satisfy the following properties:

(1) $N_F^E > 0$.

(2) (N_F^E, R_F^E) is an interior point of $\mathcal{D}$.

(3) $\overline{R}_P(\cdot,\cdot)$ is differentiable at (N_F^E, R_F^E).

(4) $\overline{R}(N)$ is differentiable at $\widetilde{N}$.

In Assumption 3, *(1)* means that flexible supply agents are willing to work for the firm under an equal treatment policy. As discussed in Lemma 4, this is equivalent to $N_P < \widetilde{N}$ (see also Proposition 12). Condition *(2)* means that it is feasible to slightly reduce or increase the number of flexible agent hours and their revenue (e.g., by prioritizing flexible supply slightly). Finally, the differentiability of $\overline{R}_P(\cdot,\cdot)$ and $\overline{R}(\cdot)$ in *(3)* and *(4)* is a natural technical assumption that is only required locally around the optimal equal treatment solution. Surprisingly, these conditions are enough to obtain an intuitive characterization of the optimality of prioritization.

Theorem 8 (Optimality of Prioritization)**.** *Suppose Assumption 1, Assumption 2 and Assumption 3 hold. Then* Profit$(\cdot)$ *is differentiable in* $R_F = R_F^E$ *and its gradient is given by:*

$$\frac{d\text{Profit}}{dR_F}(R_F^E) = \gamma\left(\frac{1}{r}\overline{R}'(\widetilde{N}) - 1\right) \tag{4.11}$$

and, therefore,

(a) if $\overline{R}'(\widetilde{N}) < r$*, there exists a policy that prioritizes private supply and has higher profit than any equal treatment policies.*

(b) If $\overline{R}'(\widetilde{N}) > r$, there exists a policy that prioritizes flexible supply and has higher profit than any equal treatment policies.

Theorem 8 proves that if $\overline{R}'(\widetilde{N}) \neq r$, then the platform should deviate from equal treatment policies and use prioritization. Remember that $\widetilde{N}$ is the total supply of the optimal equal treatment policy, and $\overline{R}(\widetilde{N})$ is the total revenue. Therefore, $\overline{R}'(\widetilde{N})$ is the marginal revenue of supply for the platform in the optimal equal treatment policy: how much extra revenue we can generate if we had one more unit of supply available. Remember that flexible supply costs r per hour in equilibrium, i.e., it is the marginal cost of increasing flexible supply. Since the private supply hours are fixed, a change in total supply hours can only come from flexible hours. The theorem then states that if the marginal revenue of adding a flexible supply hour is dominated by their marginal cost, then it is better for the firm to deter flexible supply entry by prioritizing its private supply, and vice-versa. In sum, Theorem 8 provides an intuitive and crisp characterization of the optimality of prioritization, and entails thinking about the marginal benefits and costs of incentivizing/deterring flexible supply agents by means of prioritization alone. The next subsection will formalize this intuition.

Notice that Theorem 8 establishes conditions under which prioritization is always better than equal treatment. However, our description of which supply type should be prioritized is only valid for small deviations around the optimal equal treatment policy. It is for example possible in some settings that the optimal policy prioritizes private supply even if $\overline{R}'(\widetilde{N}) > r$. Nonetheless, if $\overline{R}(\cdot)$ is concave, we can extend our result and prove our strongest result: the relationship between $\overline{R}'(\widetilde{N})$ and r uniquely determines the type of prioritization. Note that a concave maximum revenue function is a natural model for many applications, where supply has diminishing returns to scale.

Theorem 9 (Complete Characterization of Prioritization)**.** *Under the assumptions of Theorem 8, if $\overline{R}(N)$ is also strictly concave, then any optimal policies are:*

- *equal treatment policies if $\overline{R}'(\widetilde{N}) = r$,*

- *private supply prioritization policies if* $\overline{R}'(\widetilde{N}) < r$,
- *flexible supply prioritization policies if* $\overline{R}'(\widetilde{N}) > r$.

4.4.3 Characteristics of Supply Prioritization

We want to build on Theorem 9 better understand the use of prioritization policies. First, if the firm could choose the optimal pay ratio γ, it would solve Problem 4.6, with an optimal solution $(\gamma^\star, N_F^\star, R_P^\star, R_F^\star)$ (which is equal-treatment, using Theorem 7). Let $\widetilde{N}^\star \triangleq N_P + N_F^\star$ be the associated optimal total supply hours. Then Theorem 8 implies that we have $\overline{R}'(\widetilde{N}^\star) = r$. Intuitively, $\widetilde{N}^\star$ is the "optimal" level of supply, in that the marginal revenue of supply $\overline{R}'(\widetilde{N})$ is equal to the marginal cost of supply r (this is the cost of flexible supply as N_P is fixed). However, if $\gamma \neq \gamma^\star$ is fixed and not optimal, the firm does not have enough control over the flexible supply. Under equal-treatment policies, it may be *over-supplied* when $\widetilde{N} > \widetilde{N}^\star$ or *under-supplied* when $\widetilde{N} < \widetilde{N}^\star$. The following result shows that we can adapt Theorem 9 and characterize the use of prioritization policies in terms of the supply levels of the market.

Proposition 15 (Prioritization corrects supply imbalance.)**.** Under the assumptions of Theorem 9, we have:

$\overline{R}'(\widetilde{N}) = r$ if and only if $\widetilde{N} = \widetilde{N}^\star$. (perfectly-supplied market, equal treatment is optimal)

$\overline{R}'(\widetilde{N}) < r$ if and only if $\widetilde{N} > \widetilde{N}^\star$. (over-supplied market, private-prioritization is optimal)

$\overline{R}'(\widetilde{N}) > r$ if and only if $\widetilde{N} < \widetilde{N}^\star$. (under-supplied market, flexible-prioritization is optimal)

Moreover, compared to the optimal equal treatment policy, optimal private supply prioritization policies always reduce the total supply, and optimal flexible supply prioritization policies always increase the total supply.

When the market is over-supplied ($\widetilde{N} > \widetilde{N}^\star$), the firm should prioritize its private supply. This will lower the flexible supply revenue, which will decrease the total supply available in

equilibrium. In other words, we increase profit by increasing the productivity of private supply, and the associated loss of flexible supply is not important because we have too much supply. Conversely, when the market is under-supplied, we can prioritize flexible supply to increase the total supply and revenue significantly in equilibrium. As the market is under-supplied, the substantial revenue increase due to added flexible supply in equilibrium more than compensates for the loss of private supply revenue. Overall, our result states that *prioritization can increase profit by trying to restore the supply balance in the market.*

It is worth noticing that there are potential downsides related to implementing prioritization policies. First, firms may forget to anticipate equilibrium effects and always prioritize private supply. Indeed, it can be tempting for a firm that just hired employees to keep them as busy as possible instead of using contractors. This strategy would work in the short term before contractors choose to leave the market. Formally, given fixed N_P, N_F and under Assumptions 1,2,4, the firm can always increase profit by prioritizing private supply and will always decrease profit by prioritizing flexible supply. However, once the new equilibrium is reached, private supply prioritization will actually reduce profit if $\widetilde{N} < \widetilde{N}^\star$, while flexible supply prioritization will increase profit. Another potential negative impact of profit-maximizing prioritization strategies is their consequences on supply and revenue. For instance, prioritizing private agents may improve the firm's profit, but it also reduces its available supply and total revenue (affecting its service levels), which may not be desirable. We formalize this intuition in Proposition 19 in the Appendix.

So far, we have proved that, given existing private supply N_P, the firm could increase profit with prioritization policies. Nevertheless, it is not clear that this increase in profit is enough to justify operating a hybrid marketplace, as we will explore in the following section.

4.5 Optimality of Hybrid Marketplaces

This section investigates whether it is optimal for the firm to use both private and flexible supply. We already showed in Corollary 2 that using hybrid supply is not optimal if the firm can choose the optimal γ. Furthermore, even if γ is fixed, Theorem 6 shows that hybrid supply is not optimal if we limit ourselves to equal treatment policies. We will show that, with fixed γ, hybrid marketplaces can actually be optimal if we use prioritization policies. We will study the following problem, with fixed $\gamma \in (0,1)$ and optimal N_P:

$$\begin{aligned} \max_{R_P, R_F, N_F, N_P} \quad & R_P + (1-\gamma) R_F - C_P N_P \\ \text{s.t.} \quad & r N_F = \gamma R_F, \\ & (R_P, R_F) \in \mathcal{AR}(N_P, N_F). \end{aligned} \tag{4.12}$$

Our goal is to understand if $N_P > 0, N_F > 0$ can be optimal in (4.12), and we will see that this question is complex and needs careful modeling of prioritization policies.

4.5.1 Flexible supply is not needed if private supply is cheap.

When $C_p < r$, the solution to Problem (4.12) is simple:

Proposition 16 (Cheap private supply is optimal)**.** Suppose Assumption 1 holds. If $C_P < r$, all optimal solutions to Problem (4.12) verify $N_F = 0$.

The intuition behind this result is also simple. Remember that if the platform can choose γ, Lemma 5 shows that the two types of supply are "symmetric", as the firm can fully control N_P and N_F. Therefore, fixing γ imposes an additional constraint on flexible supply, making it less useful and creating potential over-supply or under-supply situations. Therefore, if $C_P < r$, private supply is cheaper and easier to control: the firm should only use it.

In the case of ride-hailing, it is because autonomous vehicles are expected to have a lower operating cost (i.e., $C_P < r$) ((Hazan et al., 2016), (Fagnant and Kockelman, 2018), (Litman,

2023)) that the industry is working on introducing this new supply source. Our proposition then shows that this has the potential of making human drivers disappear from the market.

4.5.2 The inefficiency of prioritization

In the case $C_P > r$, we will prove that it is not necessarily optimal to have $N_P = 0$. Nonetheless, the firm will need to deviate significantly from equal treatment strategies to increase profit enough to justify the higher cost of private supply. This strong prioritization may, in turn, introduce certain inefficiencies. In ride-hailing, prioritizing a given driver means that we may have to match them to arriving riders even if other de-prioritized drivers are closer. Therefore, implementing prioritization would increase the riders' average wait time compared to optimal equal treatment strategies. In turn, riders with an extended pickup time are more likely to cancel or switch to another app, and therefore prioritization would lead to revenue losses. (Krishnan et al., 2022) describes this phenomenon empirically for the ride-sharing platform Lyft. This phenomenon is also true in our model; prioritization may prevent the firm from achieving the maximum revenue attainable with its supply. Formally defining this 'prioritization inefficiency" effect is a necessary first step toward understanding the optimality of hybrid marketplaces. Let $\alpha \geqslant 0$ denote the *level of private supply prioritization*:

$$\alpha \triangleq \frac{R_P/N_P}{r/\gamma}.$$

The level of private supply prioritization measures how different the average revenue of private supply is from the average revenue of flexible supply in equilibrium. Recall that by Equation (4.1), the average revenue of flexible supply (i.e. R_F/N_F) equals r/γ. Therefore, given a feasible solution of (4.12) with $N_F > 0$, if $\alpha = 1$, the two types of supply are equally treated; if $\alpha > 1$, the private supply is prioritized; and if $\alpha < 1$, the flexible supply is prioritized. However, not all prioritization levels α are feasible given an available supply. We say the prioritization level α is achievable given the supply (N_P, N_F) if and only if there exists $R_F \geqslant 0$ such that $(\alpha r N_P/\gamma, R_F) \in \mathcal{AR}(N_P, N_F)$.

Given a supply (N_P, N_F) and a chosen feasible level of prioritization α, we now define the corresponding *efficiency loss* $\Delta R^\alpha(N_P, N_F) \geqslant 0$. It is the gap between the maximum revenue achievable without fixing α, e.g., $\overline{R}(N_P + N_F)$, and the maximum revenue achievable with α:

$$\begin{aligned} \Delta R^\alpha(N_P, N_F) \triangleq \min_{R_F \geqslant 0} \quad & \overline{R}(N_P + N_F) - \alpha r N_P/\gamma - R_F \\ \text{s.t.} \quad & (\alpha r N_P/\gamma, R_F) \in \mathcal{AR}(N_P, N_F). \end{aligned} \tag{4.13}$$

Note that the private supply revenue under prioritization α is $R_P = \alpha r N_P/\gamma$, and (4.13) finds the maximum flexible revenue R_F compatible with this prioritization. $\Delta R^\alpha(N_P, N_F)$ is the firm's revenue loss induced by a given private supply prioritization level. As we expect the inefficiency loss to scale with the amount of (de-)prioritized private supply, a quantity of interest is the ratio between the efficiency loss and the available private supply hours. We normalize this ratio with the equilibrium flexible earnings r, and define the *coefficient of prioritization inefficiency* $\beta^\alpha(N_P, N_F) \geqslant 0$:

$$\beta^\alpha(N_P, N_F) \triangleq \frac{\Delta R^\alpha(N_P, N_F)}{r N_P}$$

To better understand the meaning of α and β^α, consider an optimal equal treatment solution with supply $N_P, N_F > 0$ (e.g., an optimal solution to Problem 4.4) As we have $R_P/N_P = R_F/N_F = r/\gamma$, this corresponds to $\alpha = 1$. And because of Proposition 12, we know that the total revenue $R_P + R_F = \overline{R}(N_P + N_F)$ is maximal and therefore $\beta^1(N_P, N_F) = 0$. In summary, there is no prioritization and no efficiency loss. Suppose now that, fixing the same available supply (N_P, N_F), the firm decides to prioritize private supply and double their revenue, e.g., $R_P/N_P = 2(r/\gamma)$ and $\alpha = 2$. If we are in a situation where $\beta^2(N_P, N_F) = 0$, then there is no inefficiency and the total revenue stays equal to $r/\gamma(N_F + N_P)$: we have $R_F = r/\gamma(N_F - N_P)$. Suppose now that there is an inefficiency $\beta^2(N_P, N_F) = 1$. Then, the total revenue is now reduced by rN_P, which means that we have $R_F = r/\gamma(N_F - N_P) - rN_P$. In other words, to double the revenue of each private supply hour, we pay the equivalent of the flexible hourly earnings r in prioritization inefficiencies. In short, the notation α, β^α captures

the fact that the firms can choose various levels of prioritization, each choice corresponding to a particular inefficiency that is application-dependent.

4.5.3 Introducing expensive private supply to increase profit

We now have the tools to introduce our main result. We define Profit(N_P) to be the optimal firm profit with private supply N_P given a fixed γ, that is, the optimal objective of Problem (4.3). In the case of "expensive" private supply, $C_P > r$, we want to know when a firm should still invest in private supply and choose $N_P > 0$. This question is particularly relevant: marketplaces may wonder if they should hire employees, despite their higher costs. We again use a sensitivity analysis approach and find when Profit(dN_P) > Profit(0) for a small amount of private supply dN_P. Formally, we want to evaluate the derivative $\frac{d\text{Profit}}{dN_P}(N_P = 0)$ and understand when it is positive. We will need a few technical assumptions to make sure that $\frac{d\text{Profit}}{dN_P}(N_P = 0)$ is well defined. In particular, our prioritization parameters α and β^α must be well-defined in the limit where N_P is very small:

Definition 4. We say that α and β^α are *well-defined at* $N_P = 0$ if:

(1) There exists a neighborhood of point $(N_P = 0, N_F = \widetilde{N})$ such that α is achievable at any point in the neighborhood.

(2) $\beta^\alpha(\cdot, \cdot)$ is continuous in a neighborhood of $(0, \widetilde{N})$, and β_0^α is the limit of $\beta^\alpha(\cdot, \cdot)$ at $(0, \widetilde{N})$.

(3) The continuous extension of $\beta^\alpha(\cdot, \cdot)$ is differentiable at point $(0, \widetilde{N})$.

These conditions are purely technical and should hold in most settings. For example, without Condition *(1)*, the statement "the firm can prioritize an infinitesimal amount of private supply at level α" would not have any mathematical meaning. Condition *(2)* and *(3)* are two technical assumptions to make sure that the inefficiency coefficient is well defined for infinitely small N_P, and to ensure that $\frac{d\text{Profit}}{dN_P}(N_P = 0)$ is well defined. When α and

$\beta^\alpha(N_P, N_F)$ are well-defined at $N_P = 0$, it is mathematically meaningful to say that the firm will be able to prioritize an infinitesimal amount of supply dN_P at level α and with inefficiency coefficient β_0^α. Formally, the revenue of private supply will be $R_P = \alpha r dN_P/\gamma$ and the corresponding efficiency loss will be $\Delta R^\alpha(N_P, N_F) = r\beta_0^\alpha dN_P + o(dN_P)$.

In addition, we also assume that Assumption 1 hold and that $\overline{R}'(\widetilde{N}) < r/\gamma$. It can be easily proven that we always have$\overline{R}'(\widetilde{N}) \leqslant r/\gamma$, so we just need to avoid the "pathological" case $\overline{R}'(\widetilde{N}) = r/\gamma$. We can finally derive the closed-form of $\frac{d\text{Profit}}{dN_P}(N_P = 0)$ and characterize when marketplaces should invest in private supply:

Theorem 10 (Profit impact of introducing private supply)**.** *Given a well-defined α and β^α at $N_P = 0$, the gradient of the optimal profit in equilibrium at $N_P = 0$ can be expressed as:*

$$\frac{d\text{Profit}}{dN_P}(N_P = 0) = \left(\alpha\frac{r}{\gamma} - C_P\right) - (1-\gamma)\frac{r}{\gamma}\cdot\left(1 + \frac{(\alpha-1) + \gamma\beta_0^\alpha}{1 - \gamma\overline{R}'(\widetilde{N})/r}\right). \tag{4.14}$$

It is optimal to introduce private supply into the market with the level of prioritization α if:

$$\frac{d\text{Profit}}{dN_P}(N_P = 0) \geqslant 0 \qquad \Longleftrightarrow \qquad (\alpha-1)\frac{1 - \overline{R}'(\widetilde{N})/r}{1 - \gamma\overline{R}'(\widetilde{N})/r} \geqslant \frac{(1-\gamma)\beta_0^\alpha}{1 - \gamma\overline{R}'(\widetilde{N})/r} + \frac{C_P}{r} - 1. \tag{4.15}$$

Theorem 10 describes precisely how the firm's profit will change if it introduces private supply and chooses a prioritization level α with corresponding inefficiency β_0^α. While the expressions seem complicated, we will show that they are quite intuitive.

Consider Equation (4.14). The first term $\alpha\frac{r}{\gamma} - C_P$ is the marginal profit change due to the added private supply. Indeed, $\alpha r/\gamma$ is the revenue of private supply (note that $\frac{\partial R_P}{\partial N_P} = \alpha r/\gamma$), and C_P is the marginal cost of private supply. Correspondingly, the second term is the marginal profit change of the flexible supply. The first factor $(1-\gamma)r/\gamma$ is the average profit the firm gets from one hour of flexible supply: $1-\gamma$ is the commission rate of the platform, and r/γ is the equilibrium revenue of flexible supply). Therefore, the last term is the marginal change in flexible supply due to the introduction of private supply and the

effect of prioritization:

$$\frac{dN_F}{dN_P}(N_P = 0) = -1 - \frac{(\alpha - 1) + \gamma\beta_0^\alpha}{1 - \gamma\overline{R}'(\widetilde{N})/r}$$

The first term,‘-1”, is the replacement effect: under an equal treatment policy, each added unit of private supply replaces exactly one unit of flexible supply (as shown in Proposition 12). The second term is the effect of prioritization. The denominator is always positive, so it is a decreasing function of α and β_0^α: the more private supply prioritization and the more inefficient it is, the more we reduce the flexible supply.

If the firm uses an equal treatment policy, we have $\alpha = 1$ and $\beta_0^\alpha = 0$. Therefore, Equation (4.14) becomes $\frac{d\text{Profit}}{dN_P}(N_P = 0) = r - C_P$, and we obtain the same result as in Proposition 12.

By re-arranging the derivative of profit, expression (4.15) in the theorem establishes a condition on α, β_0^α and the parameters of the problem so that it is optimal to introduce private supply in the marketplace. The following subsection will analyze this expression to show why platforms can increase profit using hybrid supply and prioritization strategies.

4.5.4 Discussion: why is hybrid optimal?

Let us start with the special case $\beta_0^\alpha = 0$, where the firm is able to prioritize supply without any efficiency loss. Equation (4.15) becomes:

$$(\alpha - 1)\frac{1 - \overline{R}'(\widetilde{N})/r}{1 - \gamma\overline{R}'(\widetilde{N})/r} \geqslant \frac{C_P}{r} - 1. \tag{4.16}$$

The above inequality compares the benefits of a prioritization policy with level α (left-hand side) to the costs of paying for the new private supply (right-hand side). The right-hand side of Equation (4.16) is positive and represents how expensive private supply is. For example, if $C_P/r - 1 = 0.1$, then private supply is 10% more expensive than flexible supply. The left-hand side is the benefit that the firm can gain from the prioritization policy, and it must be high enough to be above the right-hand term if we want to increase profit. Specifically, $(\alpha - 1)$ describes the strength of prioritization: a high positive number is a strong private

supply prioritization, while a negative number close to -1 represents a strong flexible supply prioritization. The other term $(1 - \overline{R}'(\widetilde{N})/r)/(1 - \gamma\overline{R}'(\widetilde{N})/r)$ is a marketplace correction term. It is a function of the ratio $\overline{R}'(\widetilde{N})/r$, As we will see, it will modulate the effects of prioritization based on supply level.

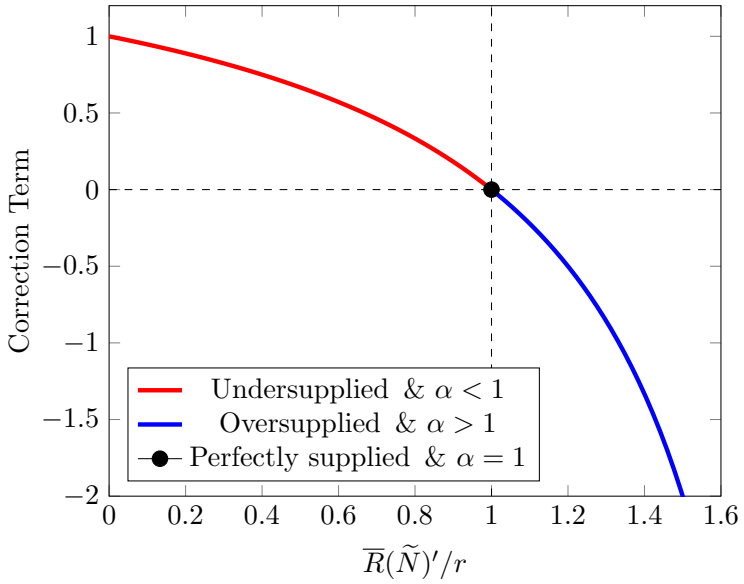


Figure 4.1: Geometrical example of the marketplace correction term with $\gamma = 0.5$.

Over-supplied market Recall from Section 4.4 that when the market is over-supplied, $\overline{R}'(\widetilde{N}) < r$, it is optimal to prioritize private supply. This is consistent with Equation (4.16). When the market is over-supplied, the correction term is positive (red curve in Figure 4.1). Therefore, the only way for Equation (4.16) to hold is to have $\alpha > 1$:we need to prioritize private supply. The correction term reaches one when $\overline{R}'(\widetilde{N})/r$ goes to zero: the more over-supplied a market is, the lesser the magnitude of prioritization needed to make hybrid supply worth it. In the best case, when $\overline{R}'(\widetilde{N}) = 0$, Equation (4.16) becomes $\alpha \geqslant C_P/r$. This is very intuitive: if private supply is twice as expensive as flexible supply ($C_P/r = 2$), we need to be able to at least double the productivity of private supply for it to be worth it: $\alpha \geqslant 2$. However, suppose the market is not too over-supplied, and $\overline{R}'(\widetilde{N})/r$ gets close to 1. In that case, the correction term goes to 0, and prioritization is much less efficient. We need a much

stronger prioritization to justify introducing private supply.

Under-supplied market If $\overline{R}'(\widetilde{N}) > r$, the market is under-supplied. The correction term (blue curve in Figure 4.1) is negative, and the only way for the left-hand side of Equation (4.16) to be positive is to have $\alpha < 1$ and prioritize flexible supply (confirming our previous results). Similarly, the more under-supplied the market is (large values of $\overline{R}'(\widetilde{N})$), the stronger the correction term is. Then, we only need to slightly prioritize flexible supply for Equation (4.16) to hold. Intuitively, when $\overline{R}'(\widetilde{N})$ is high, giving just a little extra revenue to flexible supply with prioritization can push the equilibrium significantly and add a lot of flexible supply and revenue to the market. Conversely, if we are only slightly under-supplied and $\overline{R}'(\widetilde{N})$ is close to 1, it may not be profitable to add private supply.

Perfectly supplied market The market is perfectly supplied when $\overline{R}'(\widetilde{N}) = r$, in which case the left-hand side of Equation (4.16) is zero. As $C_P > r$, it is never optimal to introduce private supply. We already had this result, as Section 4.4.2 shows that equal treatment is optimal when $\overline{R}'(\widetilde{N}) = r$, and Theorem 6 proves that hybrid is not optimal in that case.

Inefficiency loss Now that we have a complete understanding of what drives the optimality of hybrid marketplaces in the absence of inefficiencies (e.g., $\beta_0^\alpha = 0$), let us return to the general case in Equation (4.15). The inefficiency term affects the profitability of introducing private supply in two aspects. First, since β_0^α is always non-negative, the above discussion becomes necessary conditions such that it is optimal to introduce private supply. If Equation (4.16) does not hold, there is no way to make Equation (4.15) hold. After all, any loss in efficiency will reduce the revenue hurting the firm's profitability. Second, β_0^α should be an increasing function of α in practice, as a stronger prioritization (α much lower or higher than 1) should lead to a higher inefficiency. This makes it harder for the firm to use the necessary high levels of prioritization to increase profit.

In summary, Theorem 10 states that the following factors are needed for it to be profitable to introduce expensive private supply:

- The private supply cost C_P must not be too high compared to r.
- The market needs to be either significantly under-supplied or over-supplied.
- The firm needs to be able to prioritize the correct supply-type strongly enough.
- This prioritization must not be too inefficient.

Optimality of hybrid marketplaces. This section focused on showing when hybrid supply was more profitable than flexible supply only as we thought that was the most relevant question in practice. However, we did not show that hybrid supply is also better than private supply-only, and therefore we did not show that hybrid supply was optimal. Nevertheless, it turns out that it is easy to construct examples where the use of prioritization makes hybrid supply more profitable than the private-only and flexible-only alternatives:

Theorem 11 (Hybrid Marketplaces and Supply Prioritization)**.** *There exist firms for which all optimal solutions of* (4.12) *have hybrid supply. All such optimal solutions use prioritization.*

4.6 Conclusion

In this work, we study the staffing and supply management problems faced by a profit-maximizing firm that operates in a hybrid marketplace composed of private and flexible agents. We take a general approach by introducing an axiomatic framework that captures a broad range of applications. For example, private agents can be employees, rental units owned by a company (in the case of AirBnB), or autonomous vehicles. In contrast, flexible agents can be contractors, homeowners, or gig-economy workers. A critical insight is that, instead of capturing all the potential complexities associated with the firm's policies, we model the policies via the achievable revenues that supply agents can garner and then study supply prioritization in the space of achievable revenues.

		Optimal prioritization policy	Optimal Staffing
Supply of the optimal equal treatment policy	Balanced-supply (optimal commission rate)	Equal Treatment	Cheapest supply only
	Under-supplied	Flexible prioritization	Hybrid may be optimal
	Over-supplied	Private prioritization	

Table 4.1: Summary of the key results

Table 4.1 summarizes our main findings for the optimality of prioritization and staffing policies. When both types of supply are available, we find that the major benefit of a prioritization policy is to restore the supply balance. Indeed, the optimal prioritization strategy is a function of the supply imbalance induced by the optimal equal treatment policy. In particular, if the supply is perfectly balanced, there is no need to prioritize. If the market is under-supplied, it is optimal to prioritize flexible supply and increase the total supply. In contrast, if the market is over-supplied, it is optimal to prioritize private supply and reduce the size of the total supply. Based on this result, we establish that the firm should only use the cheapest supply type if the total supply is perfectly balanced. However, a hybrid marketplace may be optimal when the supply is not balanced.

We believe our model can be used as a foundation for studies that aim to explicitly uncover the role prioritization plays in hybrid marketplaces for specific settings. For instance, prioritization may take an explicit and rich form when prices, service level, and associated demand response are incorporated into a model. Finally, it would be interesting to relax our central symmetry assumption (Assumption 1) based on specific settings. For example, autonomous vehicles may be more limited and slower than human drivers, and the firm may have more control over the work of employees than independent contractors, which could lead to richer results and is another exciting direction for future work.

CHAPTER 5

Conclusion

In this thesis, we explore the aggregate-level effects of new technologies on society by examining how individual-level changes translate into broader societal impacts. Our approach combines the theories from statistics, economics, and operations research to address timely issues alongside technological advancements. Below, we summarize each chapter's findings and propose potential future research directions.

Chapter 2 presents a Bayesian model to study the societal consequences of human-AI interactions where users rationally interact with a generative AI, facing the trade-off between output fidelity and communication cost. We demonstrate that individual interactions with generative AI can lead to societal challenges. The outputs are homogenized, meaning that the AI-generated content has a lower variance than the users' original preference distribution. And this effect is amplified when AI-generated content is re-used to train the next generation of AI, potentially leading to a "homogenization death spiral." Furthermore, we also investigate the effects of AI bias, identifying who benefits or loses when using a biased AI model. Our findings show that censoring bias, which marginalizes unique preferences, negatively impacts population utility, especially for users with uncommon preferences. Directional biases, such as a slight political leaning, can influence the users' chosen output, leading to a societal bias. However, our research suggests that designing models to facilitate human-AI interactions can mitigate these risks and preserve output diversity.

In Chapter 3, we analyze the impact of introducing AVs into a fleet of HVs on ride-hailing platforms. We develop a game-theoretical queueing model where a platform aims to

maximize profit, while HVs make strategic decisions based on potential earnings compared to an outside option. Our findings indicate that incorporating AVs can degrade service levels by driving HVs out of the market. This reduction in service levels is not evenly distributed. High-demand areas are able to maintain a reasonable service level, while remote areas experience a larger decline in service level. By using a detailed simulation, we further confirm these theoretical results still hold in a more realistic setting.

Chapter 4 extends the analysis of Chapter 3 to a "hybrid marketplace" consisting of both private and flexible supply agents. We develop a general framework for supply prioritization that applies to various settings, including any firm with a mix of employees and contractors. Our main results show that, without prioritization, operating with a hybrid supply is never optimal for maximizing profit. A prioritization strategy can make it optimal for the firm to have a hybrid supply, even if the private supply is costly. In particular, prioritizing private supply can makes private agents profitable but also reduce flexible supply. Therefore, the choice of prioritization depends on market conditions. Prioritizing the private supply is favored in an "over-supplied" market, while flexible supply should be prioritized in an "under-supplied" market.

For future research, we propose several directions. First, as discussed in Chapter 2, we have demonstrated that individual rational decisions in AI interactions may lead to the societal issues of homogenization and bias. It means that the implication of utilizing AIs is not a simple cost reduction or productivity enhancement. There are a lot of potential side effects awaiting exploration and understanding. At the individual level, it is crucial to comprehend people's perceptions and attitudes towards AI, as well as how these may shift with increasing AI integration into everyday life. Empirical experiments and data analyses can shed light on these nuances. Furthermore, as suppliers increasingly incorporate AI into their services and products, it becomes important to understand how customer demand may change when products are AI-produced rather than crafted by humans. The ultimate outcome likely reflects an equilibrium between supply-side goals and demand-side

preferences. In addition, another question is how to design a better system or mechanism that fosters human-AI collaborations and avoids potential pitfalls. For a market with conflicting interests among stakeholders, it will be valuable to explore any partnership or contract that is incentive-compatible and balance the interests of all parties involved. For example, a profit-maximizing firm may actually prefer an AI that primarily focuses on frequent internet users and marginalize others to generate a higher profit. This raises the question of how a social planner can mitigate such inequalities while ensuring the firm's profitability. We believe the research presented in this thesis provides a solid foundation for addressing these complex questions and exploring these avenues further.

APPENDIX A

Human-AI Interactions and Societal Pitfalls

A.1 Characterization of Optimal Decision

In this section, we characterize the user's optimal decision. We first show the closed form of the fidelity error $e(\theta, \sigma_q)$ and illustrate how the user's decision may impact the error. Then, the optimal solution to Problem (2.4) is derived. As in Section 2.4, we assume $\mu_A = \mu_\theta$ and $\sigma_A = \sigma_\theta$.

Let's first explore how the fidelity error $e(\theta, \sigma_q)$ varies with respect to $1/\sigma_q$.

Proposition 17. For any θ, σ_q, the fidelity error is

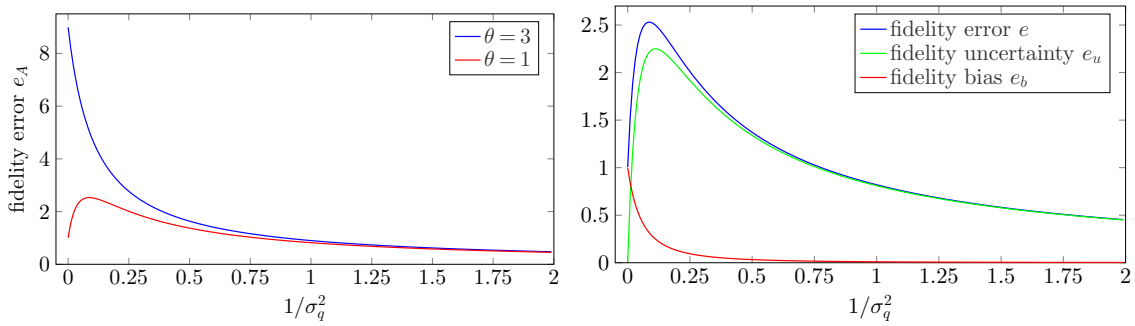
$$e(\theta, \sigma_q) = \frac{\sigma_q^2(\sigma_\theta^4 + \sigma_q^2(\mu_\theta - \theta)^2)}{(\sigma_\theta^2 + \sigma_q^2)^2} \tag{A.1}$$

Furthermore,

- $e(\theta, \sigma_q)$ increase in $d(\theta)$.
- $\lim_{\sigma_q^2 \to 0} e(\theta, \sigma_q) = 0$, $\lim_{\sigma_q^2 \to \infty} e(\theta, \sigma_q) = d(\theta)^2$
- If $d(\theta) \geq \sigma_\theta/\sqrt{2}$, $e(\theta, \sigma_q)$ is monotonically increasing in σ_q; if $d(\theta) < \sigma_\theta/\sqrt{2}$, there exists a threshold $t > 0$ such that $e(\theta, \sigma_q)$ increases in $1/\sigma_q \in (0, t)$ and decreases in $1/\sigma_q \in (t, \infty)$.

Proposition 17 reveals that for any given σ_q, the uniqueness of a user's preference results in a larger error. And if the user provides no information and simply accepts the default

output, the fidelity error increases the uniqueness metric, $d(\theta)$. Conversely, if the user offers substantial information, the fidelity error approaches zero. More intriguingly, the third part of Proposition 17 suggests that the fidelity error is monotonically decreasing in $1/\sigma_q$ only when the uniqueness $d(\theta)$ is sufficiently large (as demonstrated in the left panel of Figure A.1). In other words, if a user's preference significantly deviates from the average, any additional information to articulate their preference can yield a reduction in the AI's fidelity error. However, when a user's preference aligns closely with the majority (i.e., $d(\theta) < \sigma_\theta/\sqrt{2}$), there exists a threshold, t such that if the user is reluctant to provide sufficient information such that $1/\sigma_q > t$, sending little information may backfire, causing the user to be worse off than if they provided no information.



(a) (b)

Figure A.1: Left panel: the fidelity error with respect to $1/\sigma_q^2$. Right panel: the decomposition of fidelity error for $\theta = 1$. In both panels, We use $\mu_\theta = 0$ and $\sigma_\theta^2 = 9$.

To further investigate the cause of the non-monotonic fidelity error, as introduced in Section 2.3, we can decompose the fidelity error into a bias and a variance term,

$$e(\theta, \sigma_q) = Var(\theta_A|\theta) + [E(\theta_A|\theta) - \theta]^2,$$

We call $Var(\theta_A|\theta)$ the *fidelity uncertainty error* denoted by $e_u(\theta, \sigma_q)$, and $[E(\theta_A|\theta) - \theta]^2$ the *fidelity bias error* denoted by $e_b(\theta, \sigma_q)$. This decomposition is depicted in the right panel of Figure A.1, highlighting that the non-monotonic fidelity error is primarily driven by the

variance component. Intuitively, when a user knows that the AI's default output μ_θ is closely aligned with their preference without the need for further information, any vague information could introduce ambiguity and cause the AI to deviate from the user's true preference. For example, in Example 1, users with a neutral opinion may find it advantageous to accept the AI's default output (suppose $\mu_\theta = 0$). If they were to loosely explain their reasoning without detailing specifics, they risk introducing noisy information and receiving a less desirable result. Hence, if you're not inclined to invest enough effort in providing precise information and you're aware your preference aligns closely with the AI's default output, it may be beneficial to exert less effort or allow the AI to make decisions on your behalf. In other words, offering nothing may be preferable to providing ambiguous information.

By Proposition 17, we can solve Problem (2.4) and derive the following Lemma 6.

Lemma 6. The optimal solution to Problem (2.4) is

$$\sigma_q^\star = \begin{cases} \sqrt{\dfrac{w^\star \sigma_\theta^2}{1-w^\star}} & d(\theta) \geq \tau_d \\ \infty & \text{otherwise} \end{cases} \tag{A.2}$$

where $w^\star = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\theta-\mu_\theta)^2 - \sigma_\theta^2)}}{4((\theta-\mu_\theta)^2 - \sigma_\theta^2)}$, and $\tau_d > 0$ is a threshold that strictly increases in λ and is not less than $\sqrt{\max\{0, \sigma_\theta^2 - \dfrac{\sigma_\theta^4}{4\lambda}\}}$. In particular, $\tau_d = \dfrac{1}{2}\sigma_\theta^2 + \dfrac{1}{4}\lambda$ when $\lambda > \sigma_\theta^2$.

It is not trivial to solve Problem (2.4), since the objective function is neither concave nor convex when $d(\theta)$ is small. This non-convexity emerges from the non-monotonicity of the fidelity error, as outlined in Proposition 17. Lemma 6 implies that the users with common preferences are best suited to send no information. As discussed previously, these users may find it advantageous to rely on the AI's default output instead of introducing ambiguity.

A.1.1 Proof of the Results in Appendix A.1.

Proof. Proof of Proposition 17.

By the definition of $e(\theta, \sigma_q)$,

$$\begin{aligned} & e(\theta, \sigma_q) = E[(\theta_A(\sigma_q^2) - \theta)^2 | \theta] \\ =& E\left[\left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_q^2}(\theta + \epsilon_q) + \frac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2} \cdot \mu_\theta - \theta\right)^2 \Big| \theta\right] = \frac{\sigma_q^2(\sigma_\theta^4 + \sigma_q^2(\mu_\theta - \theta)^2)}{(\sigma_\theta^2 + \sigma_q^2)^2} \end{aligned}$$

- $\lim_{\sigma_q^2 \to 0} e(\theta, \sigma_q) = \lim_{\sigma_q^2 \to 0} \frac{\sigma_q^2(\sigma_\theta^4 + \sigma_q^2(\mu_\theta - \theta)^2)}{(\sigma_\theta^2 + \sigma_q^2)^2} = 0$. And by L'Hôpital's rule:

$$\begin{aligned} \lim_{\sigma_q^2 \to \infty} e(\theta, \sigma_q) &= \lim_{\sigma_q^2 \to \infty} \frac{\sigma_q^2(\sigma_\theta^4 + \sigma_q^2(\mu_\theta - \theta)^2)}{(\sigma_\theta^2 + \sigma_q^2)^2} \\ &= \lim_{\sigma_q^2 \to \infty} \frac{\sigma_\theta^4 + 2\sigma_q^2(\mu_\theta - \theta)^2}{2(\sigma_\theta^2 + \sigma_q^2)} \\ &= \lim_{\sigma_q^2 \to \infty} \frac{2(\mu_\theta - \theta)^2}{2} = (\mu_\theta - \theta)^2 \end{aligned}$$

- Take the derivative of $e(\theta, \sigma_q)$ with respect to σ_q^2: $\frac{\partial e(\theta, \sigma_q)}{\partial \sigma_q^2} = \frac{\partial \frac{\sigma_q^2(\sigma_\theta^4 + \sigma_q^2(\mu_\theta - \theta)^2)}{(\sigma_\theta^2 + \sigma_q^2)^2}}{\partial \sigma_q^2} = \frac{\sigma_\theta^2(\sigma_\theta^4 + \sigma_q^2(2(\mu_\theta - \theta)^2 - \sigma_\theta^2))}{(\sigma_\theta^2 + \sigma_q^2)^3}$ which is non-negative for all $\sigma_q \geq 0$ if and only if $(\mu_\theta - \theta) \geq \sigma_\theta/\sqrt{2}$. And when $(\mu_\theta - \theta) < \sigma_\theta/\sqrt{2}$, $e(\theta, \sigma_q)$ increases for

$$\sigma_q \in \left(0, \sqrt{\frac{\sigma_\theta^4}{\sigma_\theta^2 - 2(\mu_\theta - \theta)^2}}\right)$$

and decreases for

$$\sigma_q \in \left(\sqrt{\frac{\sigma_\theta^4}{\sigma_\theta^2 - 2(\mu_\theta - \theta)^2}}, \infty\right)$$

so $t = \sqrt{\frac{\sigma_\theta^2 - 2(\mu_\theta - \theta)^2}{\sigma_\theta^4}}$.

□

Proof. Proof of Lemma 6. Let $w \triangleq \frac{\sigma_q^2}{\sigma_q^2 + \sigma_\theta^2}$, and by Proposition 17, we can rewrite Problem (2.4) as:

$$w^\star(\theta) \triangleq \arg\min_{w \in [0,1]} w(1-w)\sigma_\theta^2 + w^2(\mu_\theta - \theta)^2 - \frac{\lambda}{2} \ln w \tag{A.3}$$

Let $l(w) \triangleq w(1-w)\sigma_\theta^2 + w^2(\mu_\theta - \theta)^2 - \dfrac{\lambda}{2}\ln w$. On the boundary, we have $l(0) = \infty$ and $l(1) = (\mu_\theta - \theta)^2$.

Take the first-order condition: $l'(w) = \dfrac{\partial l}{\partial w}(w) = 2((\mu_\theta - \theta)^2 - \sigma_\theta^2)w + \sigma_\theta^2 - \dfrac{\lambda}{2w} = 0$ we can get the roots of the above equation are

$$w_1 = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)},\ w_2 = \frac{-\sigma_\theta^2 - \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$$

Moreover, we have to make sure $w^\star(\theta) \in [0,1]$ and $l(w^\star(\theta)) \leq (\mu_\theta - \theta)^2$ because Problem (2.4) is non-convex.

1. $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2$

   Because $w^\star(\theta) \geq 0$ but $w_2 < 0$, it is only possible to have $w^\star(\theta) = w_1$. Also, $w_1 \leq 1 \iff -\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \leq 4((\mu_\theta - \theta)^2 - \sigma_\theta^2) \iff (\mu_\theta - \theta)^2 \geq \dfrac{1}{2}\sigma_\theta^2 + \dfrac{1}{4}\lambda$
   Additionally, when $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2$, $\dfrac{\partial l}{\partial w}$ is negative for $w < w_1$ and positive for $w > w_1$, so $l(w^\star(\theta)) < l(1) = (\mu_\theta - \theta)^2$. Therefore, when $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2$, $w^\star(\theta) = w_1$ is optimal if $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2/2 + \lambda/4$; otherwise, $w^\star(\theta) = 1$ is optimal.

2. $(\mu_\theta - \theta)^2 < \sigma_\theta^2$

   In what follows, let us discuss when we have the optimal solution $w^\star(\theta) < 1$. First, to make sure $l'(w) = 0$ has a real root (otherwise, $w^\star(\theta) = 1$ is optimal), we need $\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2) \geq 0$. That is, $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2 - \sigma_\theta^4/(4\lambda)$. In addition, we can see that $l'(w)$ is negative for $w < w_1$ or $w > w_2$. And $l'(w)$ is positive for $w \in (w_1, w_2)$ Thus, the local minimum is at $w = w_1$, and the local maximum is at $w = w_2$. This means $w = w_2$ is never optimal.

   Feasibility of $w = w_1$:

   First, because $(\mu_\theta - \theta)^2 < \sigma_\theta^2$, we must have $w_1 > 0$. Second, we want to find the

conditions such that $w_1 \leq 1$

$$w_1 = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \leq 1$$
$$\iff \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \geq 4(\mu_\theta - \theta)^2 - 3\sigma_\theta^2$$

The above inequality always holds if $(\mu_\theta - \theta)^2 \leq 3\sigma_\theta^2/4$, otherwise

$$\iff \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2) \geq (4(\mu_\theta - \theta)^2 - 3\sigma_\theta^2)^2 \iff (\mu_\theta - \theta)^2 \geq \frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\lambda$$

This implies that if $\lambda \leq \sigma_\theta^2$, we always have $w_1 \leq 1$. Thus, if $\lambda \leq \sigma_\theta^2$, we only need $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2 - \sigma_\theta^4/(4\lambda)$ such that $w = w_1 \in [0, 1]$. Otherwise, $w^\star(\theta) = 1$ is optimal. If $\lambda > \sigma_\theta^2$, we need $(\mu_\theta - \theta)^2 \geq \max\{\sigma_\theta^2/2 + \lambda/4, \sigma_\theta^2 - \sigma_\theta^4/(4\lambda)\}$. However, notice that $\sigma_\theta^2/2 + \lambda/4 \geq \sigma_\theta^2 - \sigma_\theta^4/(4\lambda)$ because $\sigma_\theta^2/2 + \lambda/4 - [\sigma_\theta^2 - \sigma_\theta^4/(4\lambda)] = (\lambda - \sigma_\theta^2)^2/(4\lambda) \geq 0$. So we need $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2/2 + \lambda/4$ such that $w = w_1 \in [0, 1]$. Otherwise, $w^\star(\theta) = 1$ is optimal.

Optimality of $w = w_1$,

Now we need to show the conditions when $w^\star(\theta) = w_1$ is optimal. Notice that $w^\star(\theta) = w_1$ is the global minimum if $w_2 \geq 1$, since $l'(w)$ is negative for $w < w_1$ and positive for $w \in (w_1, w_2)$. And,

$$w_2 = \frac{-\sigma_\theta^2 - \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \geq 1$$
$$\iff -\sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \leq 4((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2$$

The above inequality always holds if $(\mu_\theta - \theta)^2 \geq 3\sigma_\theta^2/4$, otherwise

$$\iff \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2) \geq (4((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2)^2 \iff (\mu_\theta - \theta)^2 \geq \frac{1}{2}\sigma_\theta^2 + \frac{1}{4}\lambda$$

Thus, if $\sigma_\theta^2/2 + \min\{\sigma_\theta^2, \lambda\}/4 \leq (\mu_\theta - \theta)^2 \leq \sigma_\theta^2$, $w^\star(\theta) = w_1$ is optimal.

3. Let's see what we have shown now. We have shown that if $(\mu_\theta - \theta)^2 \geq \sigma_\theta^2/2 + \lambda/4$, $w^\star(\theta) = w_1$ is feasible and optimal; if $(\mu_\theta - \theta)^2 < \sigma_\theta^2/2 + \lambda/4$ and $\lambda > \sigma_\theta^2$, $w = w_1$ is

not feasible and $w^\star(\theta) = 1$ is optimal; if $\lambda \le \sigma_\theta^2$ and $(\mu_\theta - \theta)^2 \le \sigma_\theta^2 - \sigma_\theta^4/(4\lambda)$, $w = w_1$ is not feasible and $w^\star(\theta) = 1$ is optimal. In addition, we have shown that if $\lambda \le \sigma_\theta^2$ and $(\mu_\theta - \theta)^2 \in [\sigma_\theta^2 - \sigma_\theta^4/(4\lambda),\ \sigma_\theta^2/2 + \lambda/4]$, $w = w_1 \in [0, 1]$ is feasible, but we need to show whether it is optimal since $w = 1$ is another local minimum. We want to show that if $\lambda \le \sigma_\theta^2$, there exists a threshold $\eta \ge \sigma_\theta^2 - \sigma_\theta^4/(4\lambda)$ such that when $(\mu_\theta - \theta)^2 > \eta$, $w^\star(\theta) = w_1$ is optimal; otherwise, $w^\star(\theta) = 1$ is optimal. Because we have seen that $w^\star(\theta) = w_1$ is optimal whenever $(\mu_\theta - \theta)^2 \ge \sigma_\theta^2/2 + \lambda/4$, we only need to consider the case when $(\mu_\theta - \theta)^2 \in [\sigma_\theta^2 - \sigma_\theta^4/(4\lambda), \sigma_\theta^2/2 + \lambda/4]$. That is, we want to show if $\lambda \le \sigma_\theta^2$ and $(\mu_\theta - \theta)^2 \in [\sigma_\theta^2 - \sigma_\theta^4/(4\lambda), \sigma_\theta^2/2 + \lambda/4]$,

$$g((\mu_\theta - \theta)^2) \triangleq l(1) - l(w_1) = (\mu_\theta - \theta)^2 - l\left(\frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}\right)$$

has at most one zero point. In particular, we want to show that $g((\mu_\theta - \theta)^2)$ is monotonically increasing for any $(\mu_\theta - \theta)^2 \in [\sigma_\theta^2 - \sigma_\theta^4/(4\lambda), \sigma_\theta^2/2 + \lambda/4]$. By Lemma 10,

$$\frac{\partial g}{\partial(\mu_\theta - \theta)^2} = \frac{8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 - \sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta)}{8((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

Let $h(\lambda) = 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 - \sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta)$ represents the numerator of $\dfrac{\partial g}{\partial(\mu_\theta - \theta)^2}$. We have $h(0) = 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 \ge 0$ and $h(\lambda = \sigma_\theta^2) = 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 + \sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\sigma_\theta^2((\mu_\theta - \theta)^2 - \sigma_\theta)$. Since $\lambda \le \sigma_\theta^2$ and $(\mu_\theta - \theta)^2 \le \sigma_\theta^2/2 + \lambda/4 \implies 8((\mu_\theta - \theta)^2 - \sigma_\theta)^2 \ge 2(\lambda - 2\sigma_\theta^2)((\mu_\theta - \theta)^2 - \sigma_\theta)$, $h(\lambda = \sigma_\theta^2) \ge 2(\sigma_\theta^2 - 2\sigma_\theta^2)((\mu_\theta - \theta)^2 - \sigma_\theta) + \sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\sigma_\theta^2((\mu_\theta - \theta)^2 - \sigma_\theta) = \sqrt{\Delta}(\sigma_\theta^2 - \sqrt{\Delta}) \ge 0$ (Since $\lambda \le \sigma_\theta^2$ and $(\mu_\theta - \theta)^2 \le \dfrac{1}{2}\sigma_\theta^2 + \dfrac{1}{4}\lambda \implies (\mu_\theta - \theta)^2 \le \sigma_\theta^2$). In addition, $\dfrac{\partial h}{\partial \lambda} = \dfrac{\sigma_\theta^2}{2\sqrt{\Delta}}4((\mu_\theta - \theta)^2 - \sigma_\theta) - 2((\mu_\theta - \theta)^2 - \sigma_\theta) = 2((\mu_\theta - \theta)^2 - \sigma_\theta)(\dfrac{\sigma_\theta^2}{\sqrt{\Delta}} - 1) \le 0$, because $\lambda \le \sigma_\theta^2$ and $(\mu_\theta - \theta)^2 \le \dfrac{1}{2}\sigma_\theta^2 + \dfrac{1}{4}\lambda \implies (\mu_\theta - \theta)^2 \le \sigma_\theta^2$. This implies $h(\lambda) \ge h(\lambda = \sigma_\theta^2) \ge 0$ for any $\lambda \le \sigma_\theta^2$, which further implies that $\dfrac{\partial g}{\partial(\mu_\theta - \theta)^2} \ge 0$. Therefore, if $\lambda \le \sigma_\theta^2$, $g((\mu_\theta - \theta)^2)$ is monotonically increasing for any $(\mu_\theta - \theta)^2 \in (\sigma_\theta^2 - \sigma_\theta^4/(4\lambda),\ \sigma_\theta^2/2 + \lambda/4)$. This implies that if $\lambda \le \sigma_\theta^2$, there exists a threshold $\eta \ge \sigma_\theta^2 - \sigma_\theta^4/(4\lambda)$ such that when $(\mu_\theta - \theta)^2 > \eta$, $w^\star(\theta) = w_1$ is optimal.

In summary, when $\lambda > \sigma_\theta^2$, then $\tau_d(\lambda) \triangleq \sqrt{\sigma_\theta^2/2 + \lambda/4}$ is a threshold such that $w^\star(\theta) = w_1$ is optimal if and only if $|\mu_\theta - \theta| \geq \tau_d(\lambda)$; and when $\lambda \leq \sigma_\theta^2$, then $\tau_d(\lambda) \triangleq \sqrt{\eta}$ is a threshold such that $w^\star(\theta) = w_1$ is optimal if and only if $|\mu_\theta - \theta| \geq \tau_d(\lambda)$. Additionally, it is clear that $\sigma_\theta^2/2+\lambda/4$ strictly increases in λ; and by Lemma 10, $\dfrac{\partial l(w_1)}{\partial \lambda} = \dfrac{3\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2)}{2\sqrt{\Delta}(-\sigma_\theta^2+\sqrt{\Delta})} - \dfrac{1}{2}\ln w_1 = \dfrac{3\lambda}{8\sqrt{\Delta}w_1} - \dfrac{1}{2}\ln w_1 > 0$ which implies $g((\mu_\theta-\theta)^2)$ strictly decreases in λ. Because we have shown $\dfrac{\partial g}{\partial(\mu_\theta-\theta)^2} \geq 0$, then must have η strictly increases in λ. These imply that $\tau_d(\lambda)$ strictly increases in λ.

Hence, the optimal solution to Problem (2.4) is

$$\sigma_q^\star = \begin{cases} \sqrt{\dfrac{w^\star(\theta)\sigma_\theta^2}{1-w^\star(\theta)}} & |\mu_\theta - \theta| \geq \tau_d(\lambda) \\ \infty & \text{otherwise} \end{cases} \tag{A.2}$$

where $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$, and $\tau_d(\lambda) > 0$ is a threshold that increases in λ and is not less than $\sqrt{\max\{0, \sigma_\theta^2 - \dfrac{\sigma_\theta^4}{4\lambda}\}}$.

□

A.2 Proof of the Main Results

A.2.1 Proof of the Results in Section 2.4.

A.2.1.1 Auxiliary lemmas

Lemma 7. Let $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2)}}{4((\mu_\theta-\theta)^2-\sigma_\theta^2)}$, then

1.
$$\frac{\partial w^\star(\theta)}{\partial(\mu_\theta-\theta)^2} = \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta^2)^2} \tag{A.4}$$

2.

$$\frac{\partial w^\star(\theta)}{\partial \lambda} = \frac{1}{2\sqrt{\Delta}} \tag{A.5}$$

3.

$$\frac{\partial w^\star(\theta)}{\partial \sigma_\theta^2} = \frac{(\sigma_\theta^2 - \sqrt{\Delta})(\mu_\theta - \theta)^2 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \tag{A.6}$$

where $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Proof. Proof of Lemma 7. Let $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Since $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} = \dfrac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$,

1.

$$\begin{aligned}
\frac{\partial w^\star(\theta)}{\partial (\mu_\theta - \theta)^2} &= \frac{\dfrac{1}{2\sqrt{\Delta}} 4\lambda \cdot 4((\mu_\theta - \theta)^2 - \sigma_\theta^2) - 4(-\sigma_\theta^2 + \sqrt{\Delta})}{16((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{2\lambda \cdot ((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2\sqrt{\Delta} - \Delta}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}
\end{aligned}$$

2.

$$\begin{aligned}
\frac{\partial w^\star(\theta)}{\partial \lambda} &= \frac{\dfrac{1}{2\sqrt{\Delta}} \cdot 4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \\
&= \frac{1}{2\sqrt{\Delta}}
\end{aligned}$$

3.

$$\begin{aligned}
\frac{\partial w^\star(\theta)}{\partial \sigma_\theta^2} &= \frac{\left(-1 + \frac{2\sigma_\theta^2 - 4\lambda}{2\sqrt{\Delta}}\right) 4((\mu_\theta - \theta)^2 - \sigma_\theta^2) + 4(-\sigma_\theta^2 + \sqrt{\Delta})}{16((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(-\sqrt{\Delta} + \sigma_\theta^2 - 2\lambda)((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta} + \Delta}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(\sigma_\theta^2 - \sqrt{\Delta})((\mu_\theta - \theta)^2 - \sigma_\theta^2) - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&+ \frac{-\sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(\sigma_\theta^2 - \sqrt{\Delta})((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2(\sigma_\theta^2 - \sqrt{\Delta}) + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \\
&= \frac{(\sigma_\theta^2 - \sqrt{\Delta})(\mu_\theta - \theta)^2 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}
\end{aligned}$$

□

Lemma 8. Let $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, then we can rewrite $e(\theta, \sigma_q)$ as

$$e(\theta, w) = w(1-w)\sigma_\theta^2 + w^2(\mu_\theta - \theta)^2 \tag{A.7}$$

In addition, if $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$,

1.
$$\frac{\partial e(\theta, w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2} \tag{A.8}$$

2.
$$\frac{\partial e(\theta, w^\star(\theta))}{\partial \lambda} = \frac{\sigma_\theta^2 + \sqrt{\Delta}}{4\sqrt{\Delta}} \tag{A.9}$$

3.
$$\frac{\partial e(\theta, w^\star(\theta))}{\partial \sigma_\theta^2} = \frac{[2(\mu_\theta - \theta)^2 - \sigma_\theta^2](-\sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} \tag{A.10}$$

where $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Proof. Proof of Lemma 8. Let $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, then $\sigma_q^2 = \dfrac{w\sigma_\theta^2}{1-w}$. Substitute $\sigma_q^{2\star} = \dfrac{w\sigma_\theta^2}{1-w}$ into Equation (A.1), then we have Equation (A.7).

Let $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$,

1.

$$\frac{\partial e(\theta, w^\star(\theta))}{\partial(\mu_\theta - \theta)^2}$$

$$= \sigma_\theta^2(1 - 2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} + 2(\mu_\theta - \theta)^2 w^\star(\theta)\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} + w^{\star 2}$$

Substitute Equation (A.4) into the above equation

$$= \sigma_\theta^2 \cdot \frac{\sigma_\theta^2 + 2((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sqrt{\Delta}}{2((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

$$+ 2(\mu_\theta - \theta)^2 \cdot \frac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

$$+ \frac{\sigma_\theta^4 - \sigma_\theta^2\sqrt{\Delta} + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{8((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$

$$= \frac{\sigma_\theta^4 + 2\sigma_\theta^2((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta} - (\mu_\theta - \theta)^2\sigma_\theta^2 + (\mu_\theta - \theta)^2\sqrt{\Delta}}{2((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$$

$$\cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

$$+ \frac{\sigma_\theta^4 - \sigma_\theta^2\sqrt{\Delta} + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{8((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$

$$= \frac{\sigma_\theta^2 + \sqrt{\Delta}}{2} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2} + \frac{\sigma_\theta^4 - \sigma_\theta^2\sqrt{\Delta} + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{8((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$

$$= \frac{\sigma_\theta^2(\Delta - \sigma_\theta^4) - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)(\sigma_\theta^2 + \sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2} + \frac{\sigma_\theta^4\sqrt{\Delta} - \sigma_\theta^2\Delta + 2\lambda\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$

$$= \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

2.

$$\frac{\partial e(\theta, w^\star(\theta))}{\partial \lambda}$$

$$= \sigma_\theta^2(1 - 2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial \lambda} + 2(\mu_\theta - \theta)^2 w^\star(\theta)\frac{\partial w^\star(\theta)}{\partial \lambda}$$

Substitute Equation (A.5) into the above equation

$$= \sigma_\theta^2 \cdot \frac{\sigma_\theta^2 + 2((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sqrt{\Delta}}{2((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \cdot \frac{1}{2\sqrt{\Delta}}$$

$$+ 2(\mu_\theta - \theta)^2 \cdot \frac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \cdot \frac{1}{2\sqrt{\Delta}}$$

$$= \frac{\sigma_\theta^4 + 2\sigma_\theta^2((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta} - (\mu_\theta - \theta)^2\sigma_\theta^2 + (\mu_\theta - \theta)^2\sqrt{\Delta}}{2((\mu_\theta - \theta)^2 - \sigma_\theta^2)} \cdot \frac{1}{2\sqrt{\Delta}}$$

$$= \frac{\sigma_\theta^2 + \sqrt{\Delta}}{4\sqrt{\Delta}}$$

3.

$$\frac{\partial e(\theta, w^\star(\theta))}{\partial \sigma_\theta^2}$$

$$= w^\star(\theta)(1 - w^\star(\theta)) + \sigma_\theta^2(1 - 2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial \sigma_\theta^2} + 2(\mu_\theta - \theta)^2 w^\star(\theta)\frac{\partial w^\star(\theta)}{\partial \sigma_\theta^2}$$

Substitute Equation (A.6) into the above equation

$$= \frac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$$
$$\cdot \frac{4((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2 - \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} + \frac{\sigma_\theta^2 + \sqrt{\Delta}}{2}$$
$$\cdot \frac{(\sigma_\theta^2 - \sqrt{\Delta})(\mu_\theta - \theta)^2 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)(-\sigma_\theta^2 + \sqrt{\Delta}) - (\sigma_\theta^2 - \sqrt{\Delta})^2}{16((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$+ \frac{-4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)(\mu_\theta - \theta)^2 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)(\sigma_\theta^2 + \sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{2((\mu_\theta - \theta)^2 - \sigma_\theta^2)(-\sigma_\theta^2 + \sqrt{\Delta})\sqrt{\Delta} - \sqrt{\Delta}(\sigma_\theta^4 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2) - \sigma_\theta^2\sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$+ \frac{-4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)(\mu_\theta - \theta)^2 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)(\sigma_\theta^2 + \sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{2((\mu_\theta - \theta)^2 - \sigma_\theta^2)[(-\sigma_\theta^2 + \sqrt{\Delta})\sqrt{\Delta} - \lambda\sqrt{\Delta} - 2\lambda(\mu_\theta - \theta)^2 + \lambda(\sigma_\theta^2 + \sqrt{\Delta})]}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$+ \frac{-\sqrt{\Delta}\sigma_\theta^2(\sigma_\theta^2 - \sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{2((\mu_\theta - \theta)^2 - \sigma_\theta^2)\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta}) - 2((\mu_\theta - \theta)^2 - \sigma_\theta^2)\lambda[2(\mu_\theta - \theta)^2 - \sigma_\theta^2]}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$+ \frac{\sqrt{\Delta}\sigma_\theta^2(-\sigma_\theta^2 + \sqrt{\Delta})}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{[2(\mu_\theta - \theta)^2 - \sigma_\theta^2]\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta}) - 2((\mu_\theta - \theta)^2 - \sigma_\theta^2)\lambda[2(\mu_\theta - \theta)^2 - \sigma_\theta^2]}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$
$$= \frac{[2(\mu_\theta - \theta)^2 - \sigma_\theta^2](-\sigma_\theta^2\sqrt{\Delta} + \sigma_\theta^4 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$

□

Lemma 9. Let $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, then we can rewrite $I(\sigma_q)$ as

$$I(w) = -\frac{1}{2}\ln w \tag{A.11}$$

In addition, if $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$,

1.
$$\frac{\partial I(w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = -\frac{1}{2}\cdot\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)(-\sigma_\theta^2 + \sqrt{\Delta})} \tag{A.12}$$

2.
$$\frac{\partial I(w^\star(\theta))}{\partial \lambda} = -\frac{(\mu_\theta - \theta)^2 - \sigma_\theta^2}{\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} \tag{A.13}$$

where $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

Proof. Proof of Lemma 9. Let $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$,

1.
$$\frac{\partial I(w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = -\frac{1}{2}\cdot\frac{1}{w^\star(\theta)}\cdot\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2}$$

Substitute Equation (A.4) into the above equation

$$= -\frac{1}{2}\cdot\frac{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{-\sigma_\theta^2 + \sqrt{\Delta}}\cdot\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$
$$= -\frac{1}{2}\cdot\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)(-\sigma_\theta^2 + \sqrt{\Delta})}$$

2.
$$\frac{\partial I(w^\star(\theta))}{\partial \lambda} = -\frac{1}{2}\cdot\frac{1}{w^\star(\theta)}\cdot\frac{\partial w^\star(\theta)}{\partial \lambda}$$

Substitute Equation (A.5) into the above equation

$$= -\frac{1}{2}\cdot\frac{1}{w^\star(\theta)}\cdot\frac{1}{2\sqrt{\Delta}}$$
$$= -\frac{(\mu_\theta - \theta)^2 - \sigma_\theta^2}{\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})}$$

□

Lemma 10. Let $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$, then we can rewrite Equation (2.3) as

$$l(\theta, w) = w(1-w)\sigma_\theta^2 + w^2(\mu_\theta - \theta)^2 - \frac{\lambda}{2}\ln w \tag{A.14}$$

In addition, if $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta-\theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta-\theta)^2 - \sigma_\theta^2)}$,

1.
$$\frac{\partial l(\theta, w^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = -\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{8((\mu_\theta-\theta)^2 - \sigma_\theta)^2} \tag{A.15}$$

2.
$$\frac{\partial l(\theta, w^\star(\theta))}{\partial \lambda} = \frac{3\lambda((\mu_\theta-\theta)^2 - \sigma_\theta^2)}{2\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} - \frac{1}{2}\ln w^\star(\theta) \tag{A.16}$$

where $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta-\theta)^2 - \sigma_\theta^2)$.

Proof. Proof of Lemma 10. By Lemma 8 and Lemma 9, it is clear that $l(\theta, w^\star(\theta)) = e(\theta, w^\star(\theta)) + \lambda I(w^\star(\theta)) = w^\star(\theta)(1-w^\star(\theta))\sigma_\theta^2 + w^\star(\theta)^2(\mu_\theta-\theta)^2 - \dfrac{\lambda}{2}\ln w^\star(\theta)$. In addition,

1.
$$\begin{aligned}
&\frac{\partial l(\theta, w^\star(\theta))}{\partial(\mu_\theta-\theta)^2} \\
&= \frac{\partial e(\theta, w^\star(\theta))}{\partial(\mu_\theta-\theta)^2} + \lambda\cdot\frac{\partial I(w^\star(\theta))}{\partial(\mu_\theta-\theta)^2} \\
&= \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta)^2} - \frac{\lambda}{2}\cdot\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta)(-\sigma_\theta^2+\sqrt{\Delta})} \\
&= \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2))(-\sigma_\theta^2+\sqrt{\Delta}) - 4\lambda((\mu_\theta-\theta)^2-\sigma_\theta)}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta)^2(-\sigma_\theta^2+\sqrt{\Delta})} \\
&\cdot(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2)) \\
&= \frac{(-\sigma_\theta^4 + \sigma_\theta^2\sqrt{\Delta} - 4\lambda((\mu_\theta-\theta)^2-\sigma_\theta))(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta)^2(-\sigma_\theta^2+\sqrt{\Delta})} \\
&= \frac{\sqrt{\Delta}(-\sqrt{\Delta}+\sigma_\theta^2)(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta-\theta)^2-\sigma_\theta)^2(-\sigma_\theta^2+\sqrt{\Delta})} \\
&= -\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2)}{8((\mu_\theta-\theta)^2-\sigma_\theta)^2}
\end{aligned}$$

2.

$$\begin{aligned}\frac{\partial l(\theta, w^\star(\theta))}{\partial \lambda} &= \frac{\partial e(\theta, w^\star(\theta))}{\partial \lambda} + \lambda \cdot \frac{\partial I(w^\star(\theta))}{\partial \lambda} - \frac{1}{2}\ln w^\star(\theta) \\ &= \frac{\sigma_\theta^2 + \sqrt{\Delta}}{4\sqrt{\Delta}} + \frac{\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} - \frac{1}{2}\ln w^\star(\theta) \\ &= \frac{\Delta - \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} - \frac{1}{2}\ln w^\star(\theta) \\ &= \frac{3\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{2\sqrt{\Delta}(-\sigma_\theta^2 + \sqrt{\Delta})} - \frac{1}{2}\ln w^\star(\theta)\end{aligned}$$

□

A.2.1.2 Proof of the results.

Proof. Proof of Proposition 1. Because $d(\theta) = |\mu_\theta - \theta|$ by definition and $|\mu_\theta - \theta|$ increases with $(\mu_\theta - \theta)^2$, we only have to show the change of $l(\theta, \sigma_q^\star(\theta)), I(\sigma_q^\star(\theta))$ and $e(\theta, \sigma_q^\star(\theta))$ with respect to $(\mu_\theta - \theta)^2$. By Lemma 6,

$$l(\theta, \sigma_q^\star(\theta)) = \begin{cases} l\left(\theta, \sqrt{\dfrac{w^\star(\theta)\sigma_\theta^2}{1 - w^\star(\theta)}}\right) & |\mu_\theta - \theta| \geq \tau_d(\lambda) \\ (\mu_\theta - \theta)^2 & \text{otherwise} \end{cases}$$

$$I(\sigma_q^\star(\theta)) = \begin{cases} I\left(\sqrt{\dfrac{w^\star(\theta)\sigma_\theta^2}{1 - w^\star(\theta)}}\right) & |\mu_\theta - \theta| \geq \tau_d(\lambda) \\ 0 & \text{otherwise} \end{cases}$$

$$e(\theta, \sigma_q^\star(\theta)) = \begin{cases} e\left(\theta, \sqrt{\dfrac{w^\star(\theta)\sigma_\theta^2}{1 - w^\star(\theta)}}\right) & |\mu_\theta - \theta| \geq \tau_d(\lambda) \\ (\mu_\theta - \theta)^2 & \text{otherwise} \end{cases}$$

where $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$, and $\tau_d(\lambda) > 0$ is a threshold that increases in λ and is not less than $\sigma_\theta^2 - \dfrac{\sigma_\theta^4}{4\lambda}$. Now, let's apply the results of Lemma 8, Lemma 9, and Lemma 10.

1. When $|\mu_\theta - \theta| \geq \tau_d(\lambda)$, by Lemma 10,

$$\frac{\partial l(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = -\frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{8((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

where $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$.

We only need to show $\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2) \leq 0$. When $|\mu_\theta - \theta| \geq \tau_d(\lambda)$, by the proof of Lemma 6, we know $\Delta \geq 0$, so $\sigma_\theta^4 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta) \geq 0$. Thus,

$$\begin{aligned}
&\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2) \leq 0 \qquad \text{(A.17)}\\
\iff\ &\sigma_\theta^4\Delta \leq [\sigma_\theta^4 + 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta)]^2\\
\iff\ &\sigma_\theta^4(\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)) \leq \sigma_\theta^8 + 4\lambda\sigma_\theta^4((\mu_\theta - \theta)^2 - \sigma_\theta) + 4\lambda^2((\mu_\theta - \theta)^2 - \sigma_\theta)^2\\
\iff\ &4\lambda^2((\mu_\theta - \theta)^2 - \sigma_\theta)^2 \geq 0
\end{aligned}$$

When $|\mu_\theta - \theta| < \tau_d(\lambda)$, $l(\theta, \sigma_q^\star(\theta)) = (\mu_\theta - \theta)^2 \implies \dfrac{\partial l(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = 1$. And $l(\theta, \sigma_q^\star(\theta))$ is continuous at $|\mu_\theta - \theta| = \tau_d(\lambda)$. Thus, $l(\theta, \sigma_q^\star(\theta))$ increases in $|\mu_\theta - \theta|$. By definition, $l^\star = \min(\Gamma,\ l(\theta, \sigma_q^\star(\theta))$, so $l^\star$ increases in $|\mu_\theta - \theta|$.

2. When $|\mu_\theta - \theta| \geq \tau_d(\lambda)$, by Lemma 9,

$$\frac{\partial I(\sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = -\frac{1}{2} \cdot \frac{\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)(-\sigma_\theta^2 + \sqrt{\Delta})}$$

By the proof of Lemma 6, we know $w^\star(\theta) \geq 0$ when $|\mu_\theta - \theta| \geq \tau_d(\lambda)$, so the denominator $((\mu_\theta - \theta)^2 - \sigma_\theta)(-\sigma_\theta^2 + \sqrt{\Delta}) \geq 0$. Because of the above Inequality (A.17), then $\dfrac{\partial I(\sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} \geq 0$.

When $|\mu_\theta - \theta| < \tau_d(\lambda)$, $I(\sigma_q^\star(\theta)) = 0 \implies \dfrac{\partial I(\sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = 0$.

We conclude that $I(\sigma_q^\star(\theta))$ increases in $|\mu_\theta - \theta|$.

3. Firstly, notice that $l(\theta, \sigma_q^\star) = 0$ for $d(\theta) = 0$ and we have shown that $l(\theta, \sigma_q^\star(\theta))$ monotonically increases in $d(\theta)$ in part 1. In addition, we can see that $w^\star(\theta) \to 0$ as $d(\theta) \to \infty$, which leads to $I(\sigma_q^\star(\theta)) \to \infty$ and $l(\theta, \sigma_q^\star) \to \infty$ as $d(\theta) \to \infty$. These imply

that for any $\Gamma > 0$, there must exist a threshold $\tau_a > 0$ such that $d(\theta) \leq \tau_a \iff l(\theta, \sigma_q^\star) \leq \Gamma$.

4. When $|\mu_\theta - \theta| < \tau_d(\lambda)$, by Lemma 6, $\sigma_q^\star(\theta) = \infty$, thereby $e(\theta, \sigma_q^\star(\theta)) = (\mu_\theta - \theta)^2$ and $\dfrac{\partial e(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = 1 > 0$.

   When $|\mu_\theta - \theta| \geq \tau_d(\lambda)$, by Lemma 8,

$$\frac{\partial e(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = \frac{\sigma_\theta^2(\sigma_\theta^2\sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2))}{8\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta)^2}$$

   Because of Inequality (A.17), we have $\dfrac{\partial e(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} \leq 0$.

   We conclude that if $|\mu_\theta - \theta| < \tau_d(\lambda)$, $e(\theta, \sigma_q^\star(\theta))$ increases in $(\mu_\theta - \theta)^2$; if $|\mu_\theta - \theta| \geq \tau_d(\lambda)$, $e(\theta, \sigma_q^\star(\theta))$ decreases in $|\mu_\theta - \theta|$.

□

Proof. Proof of Proposition 2. By definition, if $d(\theta) \geq \tau_a$, users will work on their own and $\theta^\star = \theta$, so $|E[\theta^\star|\theta] - \mu_\theta| = |\theta - \mu_\theta|$.

If $d(\theta) < \tau_a$, $\theta^\star = \theta_A^\star$. By Equation (2.1), we know $E[\theta_A|\theta] = \dfrac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_q^2} \cdot \theta + \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2} \cdot \mu_\theta$, so

$$|E[\theta_A^\star|\theta] - \mu_\theta| = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_q^{\star 2}(\theta)}|\theta - \mu_\theta|$$

which is 0 if $\theta = \mu_\theta$.

Additionally, since $l(\theta, \sigma_q) \to \infty$ as $\sigma_q \to 0$ and $\sigma_q = \infty$ is feasible, we must have $\sigma_q^\star(\theta) > 0$. Thus, $|E[\theta_A^\star|\theta] - \mu_\theta| < |\theta - \mu_\theta|$ whenever $\theta \neq \mu_\theta$.

□

Proof. Proof of Theorem 1. By Lemma 6, the AI's output $\theta_A(\sigma_q^\star(\sigma_\theta))$ is

$$\theta_A(\sigma_q^\star(\theta)) = \begin{cases} (1 - w^\star(\theta))q + w^\star(\theta)\mu_\theta & |\mu_\theta - \theta| \geq \tau_d(\lambda) \\ \mu_\theta & \text{otherwise} \end{cases}$$

where $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$, and $\tau_d(\lambda) > 0$ is a threshold that increases in λ and is not less than $\sigma_\theta^2 - \dfrac{\sigma_\theta^4}{4\lambda}$.

By definition, the unconditional variance of $\theta^\star$ is

$$Var(\theta^\star) = E[(\theta^\star - E[\theta^\star])^2]$$

Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative density function of $N(0,1)$, respectively. We know

$$E[\theta^\star] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \theta^\star \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

First, when $\tau_d > \tau_a$, we know that for any $\theta < \tau_a < \tau_d$, $w^\star(\theta) = 1$ and $\theta^\star = \mu_\theta$; for any $\theta > \tau_a$, $\theta^\star = \theta$, so

$$\begin{aligned}
&E[\theta^\star] \\
=&\int_{d(\theta)<\tau_a}\int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta \\
+&\int_{d(\theta)>\tau_a}\int_{-\infty}^{\infty} \theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta \\
=&\int_{d(\theta)<\tau_a} \mu_\theta \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta \\
=&\mu_\theta
\end{aligned}$$

Because $\displaystyle\int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$

$$= \int_{d(\theta)>\tau_a} (\theta - \mu_\theta) \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \mu_\theta \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

and $\displaystyle\int_{d(\theta)>\tau_a} (\theta - \mu_\theta) \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta = 0$ due to the symmetry

When $\tau_d \leq \tau_a$,

$$
\begin{aligned}
& E[\theta^\star] \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} [(1-w^\star(\theta))q + w^\star(\theta)\mu_\theta]\phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
& + \int_{d(\theta)<\tau_d} \int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
& + \int_{d(\theta)>\tau_a} \int_{-\infty}^{\infty} \theta \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} [(1-w^\star(\theta))\theta + w^\star(\theta)\mu_\theta]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)<\tau_d} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
& + \int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
& \text{Because } \int_{d(\theta)>\tau_a} \theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta = \int_{d(\theta)>\tau_a} (\theta-\mu_\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)>\tau_a} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta, \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} [(1-w^\star(\theta))(\theta-\mu_\theta) + \mu_\theta]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{d(\theta)<\tau_d} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
& + \int_{d(\theta)>\tau_a} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \int_{d(\theta)\in(\tau_d,\tau_a)} (1-w^\star(\theta))(\theta-\mu_\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
& \text{Notice that } \int_{d(\theta)\in(\tau_d,\tau_a)} (1-w^\star(\theta))(\theta-\mu_\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta = 0, \\
& \text{because } (1-w^\star(\theta))(\theta-\mu_\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) \text{ is symmetric with respect to } \theta = \mu_\theta. \\
= & \int_{-\infty}^{\infty} \mu_\theta \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
= & \mu_\theta
\end{aligned}
$$

Thus, when $\tau_d > \tau_a$,

$$
\begin{aligned}
Var(\theta^\star) &= \int_{d(\theta)>\tau_a} \int_{-\infty}^{\infty} (\mu_\theta-\theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)>\tau_a} (\mu_\theta-\theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta
\end{aligned} \tag{A.18}
$$

When $\tau_d \leq \tau_a$

$$
\begin{aligned}
& Var(\theta^\star) \\
&= \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} [\theta_A(\sigma_q^\star(\theta)) - \mu_\theta]^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)>\tau_a} \int_{-\infty}^{\infty} (\mu_\theta - \theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} [(1-w^\star(\theta))q + w^\star(\theta)\mu_\theta - \mu_\theta]^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} (1-w^\star(\theta))^2 (q-\mu_\theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} (1-w^\star(\theta))^2 (\theta+\epsilon_q-\mu_\theta)^2 \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)\in(\tau_d,\tau_a)} \int_{-\infty}^{\infty} (1-w^\star(\theta))^2 [\epsilon_q^2 - 2\epsilon_q(\mu_\theta-\theta) + (\mu_\theta-\theta)^2] \phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right) d\epsilon_q \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)\in(\tau_d,\tau_a)} (1-w^\star(\theta))^2 [\sigma_q^\star(\theta)^2 + (\mu_\theta-\theta)^2] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&= \int_{d(\theta)\in(\tau_d,\tau_a)} [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \\
&+ \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta
\end{aligned}
$$

Thus,

$$
\begin{aligned}
Var(\theta^\star) = 2\Big[& \int_{\mu_\theta+\tau_d}^{\tau_a} [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 \\
& + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2] \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta + \int_{\mu_\theta+\tau_a}^{\infty} (\mu_\theta-\theta)^2 \phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \Big]
\end{aligned} \tag{A.19}
$$

1. Now, let us first show that when $\Gamma \to \infty$, $Var(\theta^\star)$ is strictly decreasing in λ. In this case,

$$Var(\theta^\star) = 2\int_{\mu_\theta+\tau_d}^{\infty} [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$$

Let $h(\theta) \triangleq [(1-w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1-w^\star(\theta))^2(\mu_\theta-\theta)^2]$, then

$$Var(\theta_A(\sigma_q^\star(\theta))) = 2\int_{\mu_\theta+\tau_d(\lambda)}^{\infty} h(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$$

By the Leibniz integral rule,

$$\begin{aligned}\frac{\partial Var(\theta_A(\sigma_q^\star(\theta)))}{\partial\lambda} &= -2h(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)|_{\theta=\mu_\theta+\tau_d(\lambda)} \cdot \frac{\partial\sqrt{\tau_d(\lambda)}}{\partial\lambda} \\ &\quad + 2\int_{\mu_\theta+\tau_d(\lambda)}^{\infty} \frac{\partial h(\theta)}{\partial\lambda}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\end{aligned}$$

Since $\dfrac{\partial\sqrt{\tau_d(\lambda)}}{\partial\lambda} > 0$ by Lemma 6, we only need to show:

$$2\int_{\mu_\theta+\tau_d(\lambda)}^{\infty} \frac{\partial h(\theta)}{\partial\lambda}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta < 0$$

Notice that

$$\begin{aligned}&2\int_{\mu_\theta+\tau_d(\lambda)}^{\infty} \frac{\partial h(\theta)}{\partial\lambda}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\ &= 2\int_{\mu_\theta+\tau_d(\lambda)}^{\infty} \frac{\partial h(\theta)}{\partial w^\star(\theta)}\cdot\frac{\partial w^\star(\theta)}{\partial\lambda}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\ &= \int_{\mu_\theta+\tau_d(\lambda)}^{\infty} [(1-2w^\star(\theta))\sigma_\theta^2 + 2(w^\star(\theta)-1)(\mu_\theta-\theta)^2]\frac{1}{\sqrt{\Delta}}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\ &\quad \text{where } \Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta-\theta)^2-\sigma_\theta^2) \\ &= \int_{\mu_\theta+\tau_d(\lambda)}^{\infty} [2w^\star(\theta)((\mu_\theta-\theta)^2-\sigma_\theta^2) + \sigma_\theta^2 - 2(\mu_\theta-\theta)^2]\frac{1}{\sqrt{\Delta}}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\end{aligned}$$

Let $g(\theta) \triangleq [2w^\star(\theta)((\mu_\theta-\theta)^2-\sigma_\theta^2) + \sigma_\theta^2 - 2(\mu_\theta-\theta)^2]/\sqrt{\Delta}$, we want to show

$$\int_{\mu_\theta+\tau_d(\lambda)}^{\infty} g(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta < 0$$

(a) First, when $\lambda > \sigma_\theta^2/2$, we want to show $g(\theta) \leq 0$ for any $\theta \geq \mu_\theta + \tau_d(\lambda)$.

By Lemma 6, $\tau_d(\lambda) > \sqrt{\sigma_\theta^2 - \sigma_\theta^4/(4\lambda)}$, so $\tau_d(\lambda) > \sigma_\theta/\sqrt{2}$. This implies that for any $\theta \geq \mu_\theta + \tau_d(\lambda)$, $(\mu_\theta - \theta)^2 > \sigma_\theta^2/2$.

If $(\mu_\theta - \theta)^2 > \sigma_\theta^2$, $2w^\star(\theta)((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2 - 2(\mu_\theta - \theta)^2 \leq -\sigma_\theta^2 < 0$, because $w^\star(\theta) \leq 1$. And if $\frac{\sigma_\theta^2}{2} < (\mu_\theta - \theta)^2 \leq \sigma_\theta^2$, $2w^\star(\theta)((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2 - 2(\mu_\theta - \theta)^2 \leq \sigma_\theta^2 - 2(\mu_\theta - \theta)^2 < 0$, because $w^\star(\theta) > 0$. Thus, $(\mu_\theta - \theta)^2 > \frac{\sigma_\theta^2}{2}$ implies $2w^\star(\theta)((\mu_\theta - \theta)^2 - \sigma_\theta^2) + \sigma_\theta^2 - 2(\mu_\theta - \theta)^2 < 0$, which further implies $g(\theta) < 0$.

Therefore, when $\lambda > \sigma_\theta^2/2$,

$$\int_{\mu_\theta + \tau_d(\lambda)}^{\infty} g(\theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta < 0$$

(b) When $\lambda \leq \sigma_\theta^2/2$:

Let $\alpha = \lambda/\sigma_\theta^2$ (so $\lambda \leq \sigma_\theta^2/2$ implies $\alpha \leq 1/2$).

$$\begin{aligned}\Delta =& \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2) = \sigma_\theta^4\left[1 + \frac{4\lambda}{\sigma_\theta^2}\left(\frac{(\mu_\theta - \theta)^2}{\sigma_\theta^2} - 1\right)\right] \\ =& \sigma_\theta^4\left[1 + 4\alpha\left(\left(\frac{\mu_\theta - \theta}{\sigma_\theta}\right)^2 - 1\right)\right]\end{aligned}$$

Similarly, we can get

$$w^\star(\theta) = \frac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)} = \frac{-1 + \sqrt{1 + 4\alpha\left(\left(\frac{\mu_\theta - \theta}{\sigma_\theta}\right)^2 - 1\right)}}{4\left[\left(\frac{\mu_\theta - \theta}{\sigma_\theta}\right)^2 - 1\right]}$$

The substitution $x \triangleq \dfrac{\theta - \mu_\theta}{\sigma_\theta}$ yields

$$\begin{aligned}
&\int_{\mu_\theta+\tau_d(\lambda)}^{\infty} g(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{\frac{\tau_d(\lambda)}{\sigma_\theta}}^{\infty} [(1-2\hat{w}(x,\alpha))\sigma_\theta^2 + 2(\hat{w}(x,\alpha)-1)\sigma_\theta^2 x^2]\frac{1}{\sigma_\theta^2\sqrt{\hat{\Delta}(x,\alpha)}}\phi(x)\sigma_\theta dx \\
&= \int_{\frac{\tau_d(\lambda)}{\sigma_\theta}}^{\infty} [(1-2\hat{w}(x,\alpha)) + 2(\hat{w}(x,\alpha)-1)x^2]\frac{1}{\sqrt{\hat{\Delta}(x,\alpha)}}\phi(x)\sigma_\theta dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{\hat{\tau_d}(\alpha)}^{\infty} [(1-2\hat{w}(x,\alpha)) + 2(\hat{w}(x,\alpha)-1)x^2]\frac{1}{\sqrt{\hat{\Delta}(x,\alpha)}}\exp\left(-\frac{x^2}{2}\right)dx
\end{aligned}$$

where $\hat{\tau_d}(\alpha) = \dfrac{\tau_d(\lambda)}{\sigma_\theta}$, $\hat{w}(x,\alpha) = \dfrac{-1+\sqrt{1+4\alpha(x^2-1)}}{4(x^2-1)}$ and $\hat{\Delta}(x,\alpha) = 1+4\alpha(x^2-1)$. Note that

$$(1-2\hat{w}(x,\alpha)) + 2(\hat{w}(x,\alpha)-1)x^2]\frac{1}{\sqrt{\hat{\Delta}(x,\alpha)}} = \frac{1}{2}\left[1+\frac{1-4x^2}{\sqrt{1+4\alpha(x^2-1)}}\right]$$

Define

$$G(\alpha) \triangleq \int_{\hat{\tau_d}(\alpha)}^{\infty}\left[1+\frac{1-4x^2}{\sqrt{1+4\alpha(x^2-1)}}\right]\exp\left(-\frac{x^2}{2}\right)dx$$

We want to show $\forall\alpha \in [0,1/2],\ G(\alpha) < 0$.

Let's do another change of variables: $y \triangleq x^2 - 1$, which implies $dy = 2xdx$ and $x = \sqrt{y+1}$. This yields

$$G(\alpha) = \int_{\hat{\tau_d}^2(\alpha)-1}^{\infty}\left[1-\frac{3+4y}{\sqrt{1+4\alpha y}}\right]\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy$$

Let $\omega(y,\alpha) \triangleq 1-(3+4y)/\sqrt{1+4\alpha y}$. Note that

i. If $y \geq 0$, $\omega(y,\alpha)$ is increasing α.

ii. If $y \in [-3/4, 0)$, $\omega(y,\alpha)$ is decreasing α.

iii. If $y \in [-1, -3/4)$, $\omega(y,\alpha)$ is increasing α.

Correspondingly,

i. Let

$$G_0(\alpha) \triangleq \int_0^\infty \omega(y,\alpha)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy$$

we have $G_0(\alpha) \le G_0(1/2) \le G_0(1) < -0.96$.

ii. $\hat{\tau_d}^2(\alpha) - 1 \ge -3/4 \Longleftrightarrow \hat{\tau_d}^2(\alpha) \ge 1/4$

Note that $\hat{\tau_d}^2(\alpha) = \tau_d(\lambda)/\sigma_\theta$, and by the definition of $\tau_d(\lambda)$ in the proof of Lemma 6, $\tau_d(\lambda)$ solves

$$(\tau_d^2(\lambda,\sigma_\theta) - \sigma_\theta^2)m^2 + \sigma_\theta^2 m - \frac{\lambda}{2}\ln(m) = \tau_d^2(\lambda,\sigma_\theta) - \sigma_\theta^2)$$

where $m = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda(\tau_d^2(\lambda,\sigma_\theta) - \sigma_\theta^2)}}{4(\tau_d^2(\lambda,\sigma_\theta) - \sigma_\theta^2)}$. This is equivalent to that $\hat{\tau}_d(\alpha)$ solves

$$(\hat{\tau_d}^2(\alpha) - 1)m^2 + m - \frac{\alpha}{2}\ln(m) = \hat{\tau_d}^2(\alpha)$$

where $m = \dfrac{-1 + \sqrt{1 + 4\alpha(\hat{\tau_d}^2(\alpha) - 1)}}{4(\hat{\tau_d}^2(\alpha) - 1)}$.

Thus, we can get there exists $\alpha^\star$ such that $\hat{\tau_d}^2(\alpha) \ge 1/4 \Longleftrightarrow \alpha \ge \alpha^\star$. And we can numerically compute $\alpha^\star \approx 0.13845$.

Let

$$G_1(\alpha) \triangleq \int_{\hat{\tau_d}^2(\alpha)-1}^0 \omega(y,\alpha)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy$$

Since $\omega(y,\alpha)$ is decreasing in α, we have

$$\begin{aligned} G_1(\alpha) &\le \int_{\hat{\tau_d}^2(\alpha)-1}^0 \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \\ &\le \int_{-3/4}^0 \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy \end{aligned}$$

We can numerically find

$$\int_{-3/4}^0 \omega(y,\alpha^\star)\exp\left(-\frac{y+1}{2}\right)\frac{1}{2\sqrt{y+1}}dy < 0$$

Thus, $G(\alpha) = G_0(\alpha) + G_1(\alpha) < 0$.

iii. $\hat{\tau_d}^2(\alpha) - 1 < -3/4 \iff \alpha < \alpha^\star$

$$\begin{aligned}
G_1(\alpha) &= \int_{\hat{\tau_d}^2(\alpha)-1}^{\hat{\tau_d}^2(\alpha^\star)-1} \omega(y,\alpha) \exp\left(-\frac{y+1}{2}\right) \frac{1}{2\sqrt{y+1}} dy \\
&+ \int_{\hat{\tau_d}^2(\alpha^\star)-1}^{0} \omega(y,\alpha) \exp\left(-\frac{y+1}{2}\right) \frac{1}{2\sqrt{y+1}} dy \\
&\leq \int_{\hat{\tau_d}^2(\alpha)-1}^{\hat{\tau_d}^2(\alpha^\star)-1} \omega(y,\alpha^\star) \exp\left(-\frac{y+1}{2}\right) \frac{1}{2\sqrt{y+1}} dy \\
&+ \int_{\hat{\tau_d}^2(\alpha^\star)-1}^{0} \omega(y,\alpha^\star) \exp\left(-\frac{y+1}{2}\right) \frac{1}{2\sqrt{y+1}} dy \\
&= \int_{\hat{\tau_d}^2(\alpha)-1}^{0} \omega(y,\alpha^\star) \exp\left(-\frac{y+1}{2}\right) \frac{1}{2\sqrt{y+1}} dy \\
&\leq \int_{-1}^{0} \omega(y,\alpha^\star) \exp\left(-\frac{y+1}{2}\right) \frac{1}{2\sqrt{y+1}} dy
\end{aligned}$$

We can numerically find

$$\int_{-1}^{0} \omega(y,\alpha^\star) \exp\left(-\frac{y+1}{2}\right) \frac{1}{2\sqrt{y+1}} dy < 0.817$$

Thus, $G(\alpha) = G_0(\alpha) + G_1(\alpha) < -0.96 + 0.817 < 0$.

We conclude that $\forall \alpha \in [0, 1/2],\ G(\alpha) < 0$.

Hence, $Var(\theta_A(\sigma_q^{2\star}))$ strictly decreases in λ.

2. When $\lambda = 0$, we know $\forall \theta,\ w^\star(\theta) = 0,\ \theta_A^\star = \theta$. Thus,

$$\lim_{\lambda \to 0} Var(\theta^\star) = \int_{-\infty}^{\infty} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta = Var(\theta) = \sigma_\theta^2$$

When $\lambda \to \infty$, by definition, for any θ, $l \to \infty$ if σ_q is finite, so the optimal decision is $\sigma_q^\star = +\infty$ with $l^\star = (\theta - \mu_\theta)^2$. Thus, by Equation (A.18),

$$\lim_{\lambda \to \infty} Var(\theta^\star) = 2 \int_{\mu_\theta + \tau_a}^{\infty} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

And by Proposition 1, for any $\Gamma > 0$, we must have $\tau_a > 0$, so

$$\lim_{\lambda \to \infty} Var(\theta^\star) = 2 \int_{\mu_\theta + \tau_a}^{\infty} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta < 2 \int_{\mu_\theta}^{\infty} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta = Var(\theta)$$

Hence, $Var(\theta^\star) < Var(\theta)$.

3. (see Appendix A.5) Next, we want to show $Var(\theta^\star) < Var(\theta)$ if $\lambda \geq \sigma_q^2/2$ or $\Gamma \leq \hat{\Gamma}$ or $\Gamma \geq \tilde{\Gamma}$ for some $\hat{\Gamma} > 0, \tilde{\Gamma} > 0$. Let $D \triangleq Var(\theta) - Var(\theta^\star)$

   First, when $\tau_d > \tau_a$, Equation (A.18) yields

$$D = \int_{d(\theta)>0} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta - \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

   which is positive since τ_a is positive.

   Second, when $\tau_d \leq \tau_a$, Equation (A.19) yields

$$\begin{aligned} D = &\int_{\mu_\theta}^{\mu_\theta+\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta \\ &- \int_{\mu_\theta+\tau_d}^{\mu_\theta+\tau_a} [(1 - w^\star(\theta))w^\star(\theta)\sigma_\theta^2 + (1 - w^\star(\theta))^2(\mu_\theta - \theta)^2] \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta \end{aligned}$$

   We can do the same change of variables as the above steps. In particular, let $y = ((\theta - \mu_\theta)/\sigma_\theta)^2 - 1$, then we have

$$\begin{aligned} D = \frac{\sigma_\theta}{\sqrt{2\pi}} \Big[&\int_{-1}^{\hat{\tau_a}^2-1} (1+y) \frac{\exp(-(y+1)/2)}{\sqrt{y+1}} d\theta \\ &- \int_{\hat{\tau_d}^2-1}^{\hat{\tau_a}^2-1} (1-\hat{w})(1+(1-\hat{w})y) \frac{\exp(-(y+1)/2)}{\sqrt{y+1}} d\theta \Big] \end{aligned}$$

   where $\hat{\tau_a} = \tau_a/\sigma_\theta, \hat{\tau_d} = \tau_d/\sigma_\theta$, $\hat{w} = (-1 + \sqrt{1 + 4\alpha y})/(4y)$ and $\alpha = \lambda/\sigma_\theta^2$.

   (a) When $\lambda \geq \sigma_\theta^2/2$, by Lemma 6, $\tau_d \geq \sqrt{\sigma_\theta^2 - \sigma_\theta^4/(4\lambda)}$, so $\hat{\tau_d} \geq 1/\sqrt{2}$.

   Let

$$f(w) \triangleq \int_{\hat{\tau_d}^2-1}^{\hat{\tau_a}^2-1} \omega(w, y) \exp\left(-\frac{y+1}{2}\right) \frac{1}{\sqrt{y+1}} d\theta$$

   where $\omega(w, y) \triangleq (1 - w)(1 + (1 - w)y)$.

   Notice that $\dfrac{\partial \omega}{\partial w} = -1 - 2(1-w)y$, which is non-positive if and only if $(1-w)y \geq -1/2$. Because $y \geq \hat{\tau_d} - 1 > -1/2$ and $\hat{w} \in [0, 1]$, this implies that $(1 - \hat{w})y \geq -1/2$ and $\dfrac{\partial \omega}{\partial \hat{w}} \leq 0$.

   Thus,

$$\max_{w \in [0,1]} f(w) = \int_{\hat{\tau_d}^2-1}^{\hat{\tau_a}^2-1} (1+y) \exp\left(-\frac{y+1}{2}\right) \frac{1}{\sqrt{y+1}} d\theta$$

So we get a lower bound of D:

$$D \geq \frac{\sigma_\theta}{\sqrt{2\pi}} \int_{-1}^{\hat{\tau_d}^2 - 1} (y+1) \frac{\exp(-(y+1)/2)}{\sqrt{y+1}} d\theta$$

And by Lemma 6, we know $\forall \lambda > 0$, we must have $\tau_d > 0$. Thus, $D > 0$.

(b) Let $\hat{\Gamma} \triangleq l^\star(\theta)\big|_{\theta=\mu_\theta+\tau_d} > 0$

When $\Gamma \leq \hat{\Gamma}$, this means $\tau_a \leq \tau_d$, by Equation (A.18), $Var(\theta^\star) = \int_{d(\theta)>\tau_a} (\mu_\theta - \theta)^2 \phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$, which is less than $Var(\theta)$, since $\tau_a > 0$ whenever $\Gamma > 0$.

(c) Let $\acute{\Gamma} \triangleq l^\star(\theta)\big|_{\theta=\mu_\theta+\sigma_\theta/\sqrt{2}} > 0$

When $\Gamma \geq \acute{\Gamma}$, then $\tau_a \geq \sigma_\theta/\sqrt{2} \implies \hat{\tau_a} \geq 1/\sqrt{2} \implies \hat{\tau_d}^2 - 1 \geq -1/2$.

Also, in part 3 (a), we have seen that if $y \geq -1/2$, $\frac{\partial \omega}{\partial w}(\hat{w}, y) \leq 0$ (since $\hat{w} \in [0,1]$). This implies that if $y \geq -1/2$, $\omega(\hat{w}, y) \leq \omega(0, y) = (1+y)$.

And if $\hat{\tau_a}^2 - 1$ increases to $\hat{\tau_a}^2 - 1 + \xi$ for any $\xi > 0$, then the change of D is

$$\delta_D = \frac{\sigma_\theta}{\sqrt{2\pi}} \left[\int_{\hat{\tau_a}^2 - 1}^{\hat{\tau_a}^2 - 1 + \xi} [(1+y) - (1-\hat{w})(1+(1-\hat{w})y)] \frac{\exp(-(y+1)/2)}{\sqrt{y+1}} d\theta \right] \geq 0$$

This means D monotonically increases in τ_a for any $\tau_a \geq \acute{\Gamma}$.

In part 1, we have proved that $D > 0$ when $\Gamma \to \infty$, meaning that $D > 0$ when $\tau_a \to \infty$. Because D is continuous in τ_a, we either have $D > 0$ whenever $\Gamma \geq \acute{\Gamma}$ (so $\tilde{\Gamma} = \acute{\Gamma}$) or there exists another threshold $\tilde{\Gamma} > \acute{\Gamma}$ such that $D > 0$ whenever $\Gamma \geq \tilde{\Gamma}$.

□

A.2.2 Proof of the Results in Section 2.5.

A.2.2.1 Auxiliary lemmas

Lemma 11. Let $p^t \triangleq \pi_A^t(0)$ and assume $\pi_A^t(-v) = \pi_A^t(v) = (1-p^t)/2$. Then, there exist $U^t(\sigma_q, p^t)$ and $L^t(\sigma_q, p^t)$ such that

$$\begin{aligned} p^{t+1} = &\frac{(1-p_0)}{2}\left[\Phi\left(\frac{U^t(\sigma_q^t(-v),p^t)+v}{\sigma_q^t(-v)}\right) - \Phi\left(\frac{L^t(\sigma_q^t(-v),p^t)+v}{\sigma_q^t(-v)}\right)\right] \\ &+ p_0\left[\Phi\left(\frac{U^t(\sigma_q^t(0),p^t)}{\sigma_q^t(0)}\right) - \Phi\left(\frac{L^t(\sigma_q^t(0),p^t)}{\sigma_q^t(0)}\right)\right] \\ &+ \frac{(1-p_0)}{2}\left[\Phi\left(\frac{U^t(\sigma_q^t(v),p^t)-v}{\sigma_q^t(v)}\right) - \Phi\left(\frac{L^t(\sigma_q^t(v),p^t)-v}{\sigma_q^t(v)}\right)\right] \end{aligned}$$

where $U^t(\sigma_q, p^t) = -L^t(\sigma_q, p^t)$ and

$$U^t(\sigma_q, p^t) \triangleq \frac{v}{2} + \frac{\sigma_q^2}{v} \cdot \log\left(\frac{p^t}{(1-p^t)} + \sqrt{\left(\frac{p^t}{(1-p^t)}\right)^2 + 3e^{-v^2/\sigma_q^2}}\right)$$

Proof. Proof of Lemma 11. By definition,

$$\begin{aligned} \mathbb{E}[(\hat{\theta}-\theta)^2|q] &= \sum_{\theta\in\Theta}(\hat{\theta}-\theta)^2\pi_A^t(\theta|q,\sigma_q) \\ &= (\hat{\theta}+v)^2\pi_A^t(-v|q,\sigma_q) + \hat{\theta}^2\pi_A^t(0|q,\sigma_q) + (\hat{\theta}-v)^2\pi_A^t(v|q,\sigma_q) \\ &= (\hat{\theta}^2+2v\hat{\theta}+v^2)\pi_A^t(-v|q,\sigma_q) + \hat{\theta}^2\pi_A^t(0|q,\sigma_q) + (\hat{\theta}^2-2v\hat{\theta}+v^2)\pi_A^t(v|q,\sigma_q) \\ &= \hat{\theta}^2 + (2v\hat{\theta}+v^2)\pi_A^t(-v|q,\sigma_q) + (-2v\hat{\theta}+v^2)\pi_A^t(v|q,\sigma_q) \\ &= \hat{\theta}^2 + 2v\hat{\theta}(\pi_A^t(-v|q,\sigma_q) - \pi_A^t(v|q,\sigma_q)) + v^2\pi_A^t(-v|q,\sigma_q) + v^2\pi_A^t(v|q,\sigma_q) \end{aligned}$$

So θ_A^t solves:

$$\min_{\hat{\theta}\in\{-v,0,v\}}\{\hat{\theta}^2 + 2v\hat{\theta}(\pi_A^t(-v|q,\sigma_q) - \pi_A^t(v|q,\sigma_q))\}$$

Hence,

$$\theta_A^t(q,\sigma_q) = v\left[\mathbf{1}\{\pi_A^t(-v|q,\sigma_q) - \pi_A^t(v|q,\sigma_q) < -1/2\} - \mathbf{1}\{\pi_A^t(-v|q,\sigma_q) - \pi_A^t(v|q,\sigma_q) > 1/2\}\right]$$

where

$$\pi_A^t(\theta|q,\sigma_q) = \frac{\phi\left(\frac{q-\theta}{\sigma_q}\right)\pi_A^t(\theta)}{\phi\left(\frac{q+v}{\sigma_q}\right)\frac{(1-p^t)}{2} + \phi\left(\frac{q}{\sigma_q}\right)p^t + \phi\left(\frac{q-v}{\sigma_q}\right)\frac{(1-p^t)}{2}}$$

and

$$\pi_A^t(-v|q,\sigma_q) - \pi_A^t(v|q,\sigma_q) = -\frac{\exp\left(\frac{2vq}{\sigma_q^2}\right)-1}{1+\exp\left(\frac{2vq}{\sigma_q^2}\right)+2\exp\left(\frac{v^2+2vq}{2\sigma_q^2}\right)\left(\frac{p^t}{1-p^t}\right)} \triangleq d^t(q,\sigma_q).$$

We can use this to identify the values that $\theta_A^t(q,\sigma_q)$ takes. First, $\theta_A^t(q,\sigma_q) = v$ if and only is $d^t(q,\sigma_q) < -1/2$. Let $x = e^{vq/\sigma_q^2}$, then

$$\begin{aligned} d^t(q,\sigma_q) < -1/2 &\Leftrightarrow \frac{x^2-1}{1+x^2+2e^{v^2/(2\sigma_q^2)}x\frac{p^t}{(1-p^t)}} > 1/2 \\ &\Leftrightarrow 2x^2 - 2 > 1 + x^2 + 2e^{v^2/(2\sigma_q^2)}x\frac{p^t}{(1-p^t)} \\ &\Leftrightarrow x^2 - 2e^{v^2/(2\sigma_q^2)}\frac{p^t}{(1-p^t)}x - 3 > 0 \end{aligned}$$

The zeros for the equation above are:

$$\frac{2e^{v^2/(2\sigma_q^2)}\frac{p^t}{(1-p^t)} \pm \sqrt{\left(2e^{v^2/(2\sigma_q^2)}\frac{p^t}{(1-p^t)}\right)^2 + 4\cdot 3}}{2} = e^{v^2/(2\sigma_q^2)}\frac{p^t}{(1-p^t)} \pm \sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{(1-p^t)}\right)^2 + 3}.$$

We have to keep the positive zero. So then we have that

$$\begin{aligned} \theta_A^t(q,\sigma_q) = 1 &\Leftrightarrow q > \frac{\sigma_q^2}{v}\cdot\log\left(e^{v^2/(2\sigma_q^2)}\frac{p^t}{(1-p^t)} + \sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{(1-p^t)}\right)^2 + 3}\right) \\ &\Leftrightarrow q > \frac{v}{2} + \frac{\sigma_q^2}{v}\cdot\log\left(\frac{p^t}{(1-p^t)} + \sqrt{\left(\frac{p^t}{(1-p^t)}\right)^2 + 3e^{-v^2/\sigma_q^2}}\right) \triangleq U^t(\sigma_q, p^t). \end{aligned}$$

Now let's consider the case $\pi_A^t(q,\sigma_q) = -v$ which happens if and only is $d^t(q,\sigma_q) > 1/2$. Let

$x = e^{vq/\sigma_q^2}$, then

$$\begin{aligned}
d^t(q,\sigma_q) > 1/2 &\Leftrightarrow -\frac{x^2-1}{1+x^2+2e^{v^2/(2\sigma_q^2)}x\frac{p^t}{(1-p^t)}} > 1/2 \\
&\Leftrightarrow \frac{x^2-1}{1+x^2+2e^{v^2/(2\sigma_q^2)}x\frac{p^t}{(1-p^t)}} < -1/2 \\
&\Leftrightarrow 2x^2-2 < -1-x^2-2e^{v^2/(2\sigma_q^2)}x\frac{p^t}{(1-p^t)} \\
&\Leftrightarrow 3x^2+2e^{v^2/(2\sigma_q^2)}\frac{p^t}{(1-p^t)}x-1<0 \\
&\Leftrightarrow x^2+2e^{v^2/(2\sigma_q^2)}\frac{p^t}{3(1-p^t)}x-1/3<0
\end{aligned}$$

The zeros for the equation above are:

$$\begin{aligned}
&\frac{-2e^{v^2/(2\sigma_q^2)}\frac{p^t}{3(1-p^t)} \pm \sqrt{\left(2e^{v^2/(2\sigma_q^2)}\frac{p^t}{3(1-p^t)}\right)^2+4/3}}{2} \\
=&-e^{v^2/(2\sigma_q^2)}\frac{p^t}{3(1-p^t)} \pm \sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{3(1-p^t)}\right)^2+1/3}
\end{aligned}$$

We have to keep the positive zero. So then we have that

$$\begin{aligned}
\pi_A^t(q,\sigma_q) = -1 &\Leftrightarrow q < \frac{\sigma_q^2}{v}\cdot\log\left(-e^{v^2/(2\sigma_q^2)}\frac{p^t}{3(1-p^t)}+\sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{3(1-p^t)}\right)^2+1/3}\right) \\
&\Leftrightarrow q < \frac{v}{2}+\frac{\sigma_q^2}{v}\cdot\log\left(-\frac{p^t}{3(1-p^t)}+\sqrt{\left(\frac{p^t}{3(1-p^t)}\right)^2+\frac{e^{-v^2/\sigma_q^2}}{3}}\right) \triangleq L^t(\sigma_q,p^t)
\end{aligned}$$

In addition, notice that

$$
\begin{aligned}
&U^t(\sigma_q, p^t) + L^t(\sigma_q, p^t) \\
=&\frac{\sigma_q^2}{v} \cdot \left[\log\left(e^{v^2/(2\sigma_q^2)} \frac{p^t}{(1-p^t)} + \sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{(1-p^t)}\right)^2 + 3} \right) \right. \\
&\left. + \log\left(-e^{v^2/(2\sigma_q^2)} \frac{p^t}{3(1-p^t)} + \sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{3(1-p^t)}\right)^2 + 1/3} \right) \right] \\
=&\frac{\sigma_q^2}{v} \log\left[\left(e^{v^2/(2\sigma_q^2)} \frac{p^t}{(1-p^t)} + \sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{(1-p^t)}\right)^2 + 3} \right) \right. \\
&\left. \cdot \left(-e^{v^2/(2\sigma_q^2)} \frac{p^t}{3(1-p^t)} + \sqrt{e^{v^2/\sigma_q^2}\left(\frac{p^t}{3(1-p^t)}\right)^2 + 1/3} \right) \right] \\
=&\frac{\sigma_q^2}{v} \log\left[\frac{e^{v^2/(2\sigma_q^2)}p^t + \sqrt{e^{v^2/\sigma_q^2}\left(p^t\right)^2 + 3(1-p^t)^2}}{1-p^t} \cdot \frac{-e^{v^2/(2\sigma_q^2)}p^t + \sqrt{e^{v^2/\sigma_q^2}\left(p^t\right)^2 + 3(1-p^t)^2}}{3(1-p^t)} \right] \\
=&\frac{\sigma_q^2}{v} \log\left[\frac{-e^{v^2/(\sigma_q^2)}(p^t)^2 + e^{v^2/\sigma_q^2}\left(p^t\right)^2 + 3(1-p^t)^2}{3(1-p^t)^2} \right] = \frac{\sigma_q^2}{v}\log(1) = 0
\end{aligned}
$$

Therefore, the prior at $t+1$ is given by

$$
\begin{aligned}
p^{t+1} &= \mathbb{P}\left(\theta_A^t(q, \sigma_q^t(\theta)) = 0\right) \\
&= \mathbb{P}\left(q \in [L^t(\sigma_q^t(\theta), p^t), U^t(\sigma_q^t(\theta), p^t)]\right) \\
&= \mathbb{E}[\Phi\left(U^t(\sigma_q^t(\theta), p^t)\right) - \Phi\left(L^t(\sigma_q^t(\theta), p^t)\right)] \\
&= \frac{(1-p_0)}{2}\left[\Phi\left(\frac{U^t(\sigma_q^t(-v), p^t) + v}{\sigma_q^t(-v)}\right) - \Phi\left(\frac{L^t(\sigma_q^t(-v), p^t) + v}{\sigma_q^t(-v)}\right)\right] \\
&+ p_0\left[\Phi\left(\frac{U^t(\sigma_q^t(0), p^t)}{\sigma_q^t(0)}\right) - \Phi\left(\frac{L^t(\sigma_q^t(0), p^t)}{\sigma_q^t(0)}\right)\right] \\
&+ \frac{(1-p_0)}{2}\left[\Phi\left(\frac{U^t(\sigma_q^t(v), p^t) - v}{\sigma_q^t(v)}\right) - \Phi\left(\frac{L^t(\sigma_q^t(v), p^t) - v}{\sigma_q^t(v)}\right)\right]
\end{aligned}
$$

□

A.2.2.2 Proof of the results.

Proof. Proof of Lemma 1. If $\sigma_q = \infty$, $\pi_A^t(\theta|q,\sigma_q) = \pi_A^t(\theta)$, so $I(\theta,\sigma_q) = 0$. In addition, suppose $\pi_A^t(-v) = \pi_A^t(v)$ for some t. If $\sigma_q = \infty$, $\theta_A^t = \arg\min_{\hat{\theta}\in\Theta} \sum_{\theta\in\Theta}(\hat{\theta}-\theta)^2 \pi_A^t(\theta|q,\sigma_q) = \arg\min_{\hat{\theta}\in\Theta} \sum_{\theta\in\Theta}(\hat{\theta}-\theta)^2 \pi_A^t(\theta) = 0$, so $e^t(0,\infty) = 0$. This means that any user with $\theta = 0$ can achieve zero utility loss if they share no information. Thus, $\sigma_q^{\star t}(0) = \infty$. On the other hand, $\forall \sigma_q,\ l^t(-v,\sigma_q) = l^t(v,\sigma_q)$ because $e^t(-v,\sigma_q) = e^t(v,\sigma_q)$. This implies $\theta^{\star t}(-v) = \theta^{\star t}(v)$, which further implies $\pi_A^{t+1}(-v) = \pi_A^{t+1}(v)$ and $\sigma_q^{\star t}(0) = \infty$. Hence, $\forall t,\ \sigma_q^{\star t}(0) = \infty$ and $\pi_A^t(-v) = \pi_A^t(v)$.

□

Proof. Proof of Corollary 1. By definition, $Var(\theta^{\star t}) = E[(\theta^{\star t} - E[\theta^{\star t}])^2]$ and $E[\theta^{\star t}] = 0$ because of Lemma 1. This means $Var(\theta^{\star t}) = E[\theta^{\star t 2}] = v^2(1-\pi_A^{t+1}(0))$. And we know $Var(\theta) = v^2(1-p_0)$, so we only need to show $\pi_A^{t+1}(0) \geq p_0$. However, this is always true because $\pi_A^{t+1}(0) = \mathbb{P}(\theta^{\star t} = 0) \geq \mathbb{P}(\theta^{\star t} = 0|\theta = 0)\mathbb{P}(\theta = 0) = 1 \cdot \mathbb{P}(\theta = 0) = p_0$ by Lemma 1. Therefore, $\forall t,\ Var(\theta^{\star t}) \leq Var(\theta)$.

Second,

$$Var(\theta^{\star t}) = Var(\theta) \iff \mathbb{P}(\theta^{\star t} = 0) = p_0 \iff \mathbb{P}(\theta^{\star t} = 0|\theta = -v) = \mathbb{P}(\theta^{\star t} = 0|\theta = v) = 0$$
$$\iff \mathbb{P}(\theta_A^{\star t} = 0|\theta = -v) = \mathbb{P}(\theta_A^{\star t} = 0|\theta = v) = 0 \iff \sigma_q^{\star t}(-v) = \sigma_q^{\star t}(v) = 0$$

□

Proof. Proof of Proposition 3. Because $\pi_A^t(-v) = \pi_A^t(v)$ and $\sigma_q^t(-v) = \sigma_q^t(v)$, we have $E[\theta_A^{t+1}] = 0$ and $Var(\theta_A^{t+1}) = v^2(1-\pi_A^{t+1}(0))$. Let $p^t(\sigma_q) \triangleq \pi_A^t(0)$. Thus, what we want to show is

1. p^{t+1} strictly increases in p^t.
2. p^{t+1} strictly increases in σ_q.

1. For the first statement, by Lemma 11, we only need to show $U^t(\sigma_q, p^t)$ strictly increases in p^t. In fact, because

$$U^t(\sigma_q, p^t) = \frac{v}{2} + \frac{\sigma_q^2}{v} \cdot \log\left(\frac{p^t}{(1-p^t)} + \sqrt{\left(\frac{p^t}{(1-p^t)}\right)^2 + 3e^{-v^2/\sigma_q^2}}\right)$$

as shown in Lemma 11, it is clear that $U^t(\sigma_q, p^t)$ strictly increases in p^t.

2. For the second statement, we want to show $\partial p^{t+1}/\partial \sigma_q > 0$. Because $\sigma_q^t(-v) = \sigma_q^t(v) = \sigma_q$, we can simplify the expression of p^{t+1} in Lemma 11:

$$\begin{aligned} p^{t+1} = p_0 &\left[\Phi\left(\frac{U^t(\sigma_q^t(0), p^t)}{\sigma_q^t(0)}\right) - \Phi\left(\frac{L^t(\sigma_q^t(0), p^t)}{\sigma_q^t(0)}\right)\right] \\ &+ (1-p_0)\left[\Phi\left(\frac{U^t(\sigma_q, p^t) - v}{\sigma_q}\right) - \Phi\left(\frac{-U^t(\sigma_q, p^t) - v}{\sigma_q}\right)\right] \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\partial p^{t+1}}{\partial \sigma_q} &\propto \phi\left(\frac{U^t - v}{\sigma_q}\right) \cdot \frac{\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v}{\sigma_q^2} - \phi\left(\frac{-U^t - v}{\sigma_q}\right) \cdot \frac{-\frac{\partial U^t}{\partial \sigma_q}\sigma_q + U^t + v}{\sigma_q^2} \\ &\propto exp\left(-\frac{(U^t - v)^2}{2\sigma_q^2}\right) \cdot \frac{\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v}{\sigma_q^2} + exp\left(\frac{-(U^t + v)^2}{2\sigma_q^2}\right) \cdot \frac{\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t - v}{\sigma_q^2} \\ &\propto exp\left(\frac{vU^t}{\sigma_q^2}\right)\left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v\right) + exp\left(-\frac{vU^t}{\sigma_q^2}\right)\left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t - v\right) \\ &\propto exp\left(\frac{2vU^t}{\sigma_q^2}\right)\left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v\right) + \left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t - v\right) \triangleq f \end{aligned}$$

We want to show $f > 0$.

Let $x \triangleq exp(v^2/(2\sigma_q^2))$ and $y \triangleq p^t/(1-p^t)$. With some algebra, we can get

$$\begin{aligned} \frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t &= U^t - v\frac{xy(\sqrt{x^2y^2+3} + xy)}{xy(\sqrt{x^2y^2+3} + xy) + 3} \\ &= U^t - v\frac{1}{1 + 3/[xy(\sqrt{x^2y^2+3} + xy)]} > -v \end{aligned}$$

where the last inequality is given by $U^t \geq 0$, $x \geq 0$ and $y \geq 0$. Therefore,

$$f > exp\left(\frac{2vU^t}{\sigma_q^2}\right)\left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v\right) - 2v$$

We want to show $exp\left(\frac{2vU^t}{\sigma_q^2}\right)\left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v\right) > 2v$.

With some algebra, we can get

$$exp\left(\frac{2vU^t}{\sigma_q^2}\right) = (xy+\sqrt{x^2y^2+3})^2 \text{ and } \frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v = U^t + \frac{3v}{xy(xy+\sqrt{x^2y^2+3})+3}$$

Because $U^t \geq 0$,

$$exp\left(\frac{2vU^t}{\sigma_q^2}\right)\left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v\right) \geq v \cdot \frac{3(xy+\sqrt{x^2y^2+3})^2}{xy(xy+\sqrt{x^2y^2+3})+3}$$

Because $x \geq 0$ and $y \geq 0$,

$$(xy+\sqrt{x^2y^2+3})^2 = x^2y^2 + 2xy\sqrt{x^2y^2+3} + x^2y^2 + 3 > xy(xy+\sqrt{x^2y^2+3}) + 3$$

Thus,

$$exp\left(\frac{2vU^t}{\sigma_q^2}\right)\left(\frac{\partial U^t}{\partial \sigma_q}\sigma_q - U^t + v\right) \geq 3v > 2v$$

Hence, we have $\partial p^{t+1}/\partial \sigma_q > 0$.

□

Proof. Proof of Theorem 2. In Lemma 1, we have seen $\sigma_q^{\star t}(0) = \infty$ and $\sigma_q^{\star t}(-v) = \sigma_q^{\star t}(v)$. Clearly, these still hold if there is another constraint such that every user cannot share more information than the previous iteration). That is, if there is another constraint $\sigma_q^{t+1}(\theta) \geq \sigma_q^t(\theta)$ for any θ, we still have $\forall t,\ \sigma_q^{\star t}(0) = \infty$ and $\sigma_q^{\star t}(-v) = \sigma_q^{\star t}(v)$. To simplify the notation, let $\sigma_q^t = \sigma_q^t(-v) = \sigma_q^t(v)$ for some σ_q^t. By Proposition 3, $Var(\theta_A^{t+1}|\sigma_q^{t+1} > \sigma_q^{\star t}) < Var(\theta_A^{t+1}|\sigma_q^{t+1} = \sigma_q^{\star t})$. Therefore, to show $Var(\theta^{\star(t+1)}) < Var(\theta^{\star t})$, we only need to show $Var(\theta_A^{t+1}|\sigma_q^{t+1} = \sigma_q^{\star t}) \leq Var(\theta^{\star t}) = Var(\theta_A^t|\sigma_q^{\star t})$, as we assume $\Gamma = \infty$. That is, we want to show the variance of outputs will not increase if the user with $\theta \neq 0$ shares the same amount of information as in the previous iteration.

By Proposition 3, we have $Var(\theta_A^{t+1}|\sigma_q^{t+1} = \sigma_q^{\star t}) \leq Var(\theta_A^t|\sigma_q^{\star t})$ is true if $Var(\theta^{\star t}) \leq Var(\theta^{\star(t-1)})$. This implies that $Var(\theta^{\star(t+1)}) < Var(\theta^{\star t})$ if $Var(\theta^{\star t}) \leq Var(\theta^{\star(t-1)})$. Furthermore, by Corollary 1, we have $Var(\theta^{\star 0}) \leq Var(\theta)$. Hence, we must have $\forall t,\ Var(\theta^{\star(t+1)}) <$

$Var(\theta^{\star t})$ by mathematical induction. This concludes that there exists a homogenization death spiral if $\Gamma = \infty$ and users cannot share more information than in the previous iteration.

Next, we want to show the existence of a counter-example when either Γ is small enough or users can share more information than in the previous iteration.

First, let $\hat{l}^t$ denote the optimal utility loss of using the AI for the unique user with $\theta \neq 0$ at time t when $\Gamma = \infty$. Let $\overline{\hat{l}^{\star t}} = \sup\{\hat{l}^{\star t}\}_{t=0}^{\infty}$ be the supremum of the utility loss over time. Suppose Γ reduces to some value strictly less than $\overline{\hat{l}^{\star t}}$, then there exists a period $\hat{t}$ such that the utility loss of using the AI is higher than Γ (i.e., $\hat{l}^{\hat{t}} > \Gamma$). Thus, at time $\hat{t}$, we must have $Var(\theta_A^{\star \hat{t}}) = Var(\theta) \geq Var(\theta_A^{\star(\hat{t}-1)})$. We conclude that the homogenization death spiral is broken if $\Gamma < \overline{\hat{l}^{\star t}}$.

On the other hand, suppose $\Gamma = \infty$. By Proposition 3, we know $Var(\theta_A^{t+1})$ strictly decreases in σ_q whenever $Var(\theta_A^{t+1}) > 0$. And by Corollary 1, $Var(\theta_A^{t+1}) = Var(\theta) \geq Var(\theta_A^t)$ if $\sigma_q^{\star t}(-v) = \sigma_q^{\star t}(v) = 0$. Because of continuity, there exists a threshold $\hat{\sigma_q} > 0$ such that $Var(\theta_A^{t+1}) \geq Var(\theta_A^t)$ if $\sigma_q^{\star t}(-v) = \sigma_q^{\star t}(v) \leq \hat{\sigma_q}$. □

A.2.3 Proof of the Results in Section 2.6.

A.2.3.1 Auxiliary lemmas

Lemma 12. For any θ, σ_q^2,

$$e(\theta, \sigma_q) = \frac{\sigma_q^2(\sigma_A^4 + \sigma_q^2(\mu_A - \theta)^2)}{(\sigma_A^2 + \sigma_q^2)^2} \tag{A.20}$$

In addition,

- Both $l(\theta, \sigma_q^2)$ and $e(\theta, \sigma_q)$ strictly increase in $(\mu_A - \theta)^2$.
- Both $l(\theta, \sigma_q^2)$ and $e(\theta, \sigma_q)$ strictly decrease in σ_A^2 for $\sigma_A^2 < (\mu_A - \theta)^2$ and increase in σ_A^2 for $\sigma_A^2 \geq (\mu_A - \theta)^2$.

Proof. Proof of Lemma 12. By Equation (2.1), $\theta_A = \frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}q + \frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}\mu_A$. Then,

$$\begin{aligned}
e(\theta,\sigma_q) &= E\left[\left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}(\theta+\epsilon_q) + \frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}\mu_A - \theta\right)^2 |\theta\right] \\
&= E\left[\left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}\epsilon_q + \frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}(\mu_A-\theta)\right)^2 |\theta\right] \\
&= \left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}\right)^2 E[\epsilon_q^2] + \left(\frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}(\mu_A-\theta)\right)^2 \\
&= \left(\frac{\sigma_A^2}{\sigma_A^2+\sigma_q^2}\right)^2 \sigma_q^2 + \left(\frac{\sigma_q^2}{\sigma_A^2+\sigma_q^2}(\mu_A-\theta)\right)^2 \\
&= \frac{\sigma_q^2(\sigma_A^4 + \sigma_q^2(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^2}
\end{aligned}$$

It is clear that $e(\theta,\sigma_q)$ strictly increases in $(\mu_A-\theta)^2$, and $l(\theta,\sigma_q^2)$ strictly increases in $(\mu_A-\theta)^2$ (Note that $I(\sigma_q^2)$ does not depend on either μ_A or σ_A).

Take the derivative of $e(\theta,\sigma_q)$ with respect to σ_A^2:

$$\begin{aligned}
\frac{\partial e(\theta,\sigma_q)}{\partial \sigma_A^2} &= \frac{2\sigma_q^2\sigma_A^2(\sigma_A^2+\sigma_q^2)^2 - 2(\sigma_A^2+\sigma_q^2)\sigma_q^2(\sigma_A^4+\sigma_q^2(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^4} \\
&= \frac{2\sigma_q^2\sigma_A^2(\sigma_A^2+\sigma_q^2) - 2\sigma_q^2(\sigma_A^4+\sigma_q^2(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^3} \\
&= \frac{2\sigma_q^4(\sigma_A^2-(\mu_A-\theta)^2)}{(\sigma_A^2+\sigma_q^2)^3}
\end{aligned}$$

Thus, $\frac{\partial e(\theta,\sigma_q)}{\partial \sigma_A^2} < 0$ if $\sigma_A^2 < (\mu_A-\theta)^2$, and $\frac{\partial e(\theta,\sigma_q)}{\partial \sigma_A^2} \geq 0$ if $\sigma_A^2 \geq (\mu_A-\theta)^2$. This implies that both $l(\theta,\sigma_q^2)$ and $e(\theta,\sigma_q)$ strictly decrease in σ_A^2 for $\sigma_A^2 < (\mu_A-\theta)^2$ and increase in σ_A^2 for $\sigma_A^2 \geq (\mu_A-\theta)^2$.

□

Lemma 13. Let $w^\star(\theta,\lambda) = \frac{\sigma_q^{\star 2}(\theta,\lambda)}{\sigma_A^2+\sigma_q^{\star 2}(\theta,\lambda)}$. $\forall \theta,\mu_A,\sigma_A,\lambda_1>\lambda_2$, $w^\star(\theta,\lambda_1) \geq w^\star(\theta,\lambda_2)$.

Proof. Proof of Lemma 13. Firstly, by Equation (A.20), we can write the objective function

as:

$$l(\theta, w, \lambda) = w(1-w)\sigma_A^2 + w^2(\mu_A - \theta)^2 - \frac{\lambda}{2}\ln w$$

where $w = \sigma_q^2/(\sigma_A^2 + \sigma_q^2) \in [0, 1]$.

For the sake of contradiction, assume $w^\star(\theta, \lambda_1) < w^\star(\theta, \lambda_2)$ for some θ. Since $I(w) = -\ln(w)/2$ strictly decreases in w, we have $\delta_I \triangleq I(w^\star(\theta, \lambda_1)) - I(w^\star(\theta, \lambda_2)) > 0$. Let $\delta_e \triangleq e(\theta, w^\star(\theta, \lambda_1)) - e(\theta, w^\star(\theta, \lambda_2))$. Because $w^\star(\theta, \lambda_1)$ is optimal when $\lambda = \lambda_1$, we must have $l(\theta, w^\star(\theta, \lambda_1), \lambda_1) - l(\theta, w^\star(\theta, \lambda_2), \lambda_1) < 0$. This implies $\delta_e < 0$ and $\delta_e + \lambda_1\delta_I < 0$. However, because $\lambda_1 > \lambda_2$ and $\delta_I > 0$, we must have $\delta_e + \lambda_2\delta_I < \delta_e + \lambda_1\delta_I < 0$, meaning that $l(\theta, w^\star(\theta, \lambda_1), \lambda_2) - l(\theta, w^\star(\theta, \lambda_2), \lambda_2) < 0$. This contradicts the assumption that $w^\star(\theta, \lambda_2)$ is optimal when $\lambda = \lambda_2$. Therefore $\forall\theta$, $w^\star(\theta, \lambda_1) \geq w^\star(\theta, \lambda_2)$ whenever $\lambda_1 > \lambda_2$.

□

A.2.3.2 Proofs of the results

Proof. Proof of Proposition 4.

- Suppose $|\mu_{A_1} - \theta| > |\mu_{A_2} - \theta|$ for some $\mu_{A_1}, \mu_{A_2}, \theta$. Let $\sigma_{q_1}^\star$ and $\sigma_{q_2}^\star$ denote the optimal decision for user θ in Problem (2.4) when $\mu_A = \mu_{A_1}$ and $\mu_A = \mu_{A_2}$, respectively. By definition of l in Equation (2.3), let $l_1^\star = l(\theta, \sigma_{q_1}^\star, \mu_{A_1})$ and $l_2^\star = l(\theta, \sigma_{q_2}^\star, \mu_{A_2})$.

  We want to show $l_1^\star > l_2^\star$. For the sake of contradiction, suppose $l_1^\star \leq l_2^\star$. By Lemma 12, $l_1^\star = l(\theta, \sigma_{q_1}^\star, \mu_{A_1}) > l(\theta, \sigma_{q_1}^\star, \mu_{A_2})$. This implies $l(\theta, \sigma_{q_1}^\star, \mu_{A_2}) < l_2^\star = l(\theta, \sigma_{q_2}^\star, \mu_{A_2})$. This contradicts the assumption that $\sigma_{q_2}^\star$ minimizes $l(\theta, \sigma_q, \mu_{A_2})$. Therefore, $l_1^\star > l_2^\star$. We conclude that $l^\star$ strictly increases in $|\mu_A - \theta|$.

- Suppose $\sigma_{A_1} < \sigma_{A_2} < |\mu_A - \theta|$ for some $\sigma_{A_1}, \sigma_{A_2}, \mu_A, \theta$. Let $\sigma_{q_1}^\star$ and $\sigma_{q_2}^\star$ denote the optimal decision for user θ in Problem (2.4) when $\sigma_A = \sigma_{A_1}$ and $\sigma_A = \sigma_{A_2}$, respectively. By definition of l in Equation (2.3), let $l_1^\star = l(\theta, \sigma_{q_1}^\star, \sigma_{A_1})$ and $l_2^\star = l(\theta, \sigma_{q_2}^\star, \sigma_{A_2})$.

  We want to show $l_1^\star > l_2^\star$. For the sake of contradiction, suppose $l_1^\star \leq l_2^\star$. By Lemma 12,

$l_1^\star = l(\theta, \sigma_{q_1}^\star, \sigma_{A_1}) > l(\theta, \sigma_{q_1}^\star, \sigma_{A_2})$. This implies $l(\theta, \sigma_{q_1}^\star, \sigma_{A_2}) < l_2^\star = l(\theta, \sigma_{q_2}^\star, \sigma_{A_2})$. This contradicts the assumption that $\sigma_{q_2}^\star$ minimizes $l(\theta, \sigma_q, \sigma_{A_2})$. Therefore, $l_1^\star > l_2^\star$. We conclude that $l^\star$ strictly decreases in σ_A when $\sigma_A < |\mu_A - \theta|$.

Similarly, when $|\mu_A - \theta| \leq \sigma_{A_1} < \sigma_{A_2}$, we want to show $l_1^\star \leq l_2^\star$. For the sake of contradiction, suppose $l_1^\star > l_2^\star$. By Lemma 12, $l_2^\star = l(\theta, \sigma_{q_2}^\star, \sigma_{A_2}) > l(\theta, \sigma_{q_2}^\star, \sigma_{A_1})$. This implies $l(\theta, \sigma_{q_2}^\star, \sigma_{A_1}) < l_1^\star = l(\theta, \sigma_{q_1}^\star, \sigma_{A_1})$. This contradicts the assumption that $\sigma_{q_1}^\star$ minimizes $l(\theta, \sigma_q, \sigma_{A_1})$. Therefore, $l_1^\star \leq l_2^\star$. We conclude that $l^\star$ strictly increases in σ_A when $\sigma_A \geq |\mu_A - \theta|$.

□

Proof. Proof of Proposition 5. Let $\phi(\cdot)$ denote the probability density function of $N(0,1)$. And let $w = \sigma_q^2/(\sigma_A^2 + \sigma_q^2)$.

- Let us first show $E[l^\star(\theta, \mu_A)]$ is minimized at $\mu_A = \mu_\theta$. That is, $\forall \mu_{A1} \neq \mu_\theta$, we want to show $E[l^\star(\theta, \mu_{A1})] > E[l^\star(\theta, \mu_\theta)]$. Without loss of generality, suppose $\mu_{A1} > \mu_\theta$.

  By definition,

  $$E[l^\star(\theta, \mu_A)] = \int_{-\infty}^{\infty} l^\star(\theta, \mu_A)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

  So we want to show

  $$\int_{-\infty}^{\infty} [l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta > 0$$

  By Equation (A.20), $\forall \sigma_q, \theta_1, \theta_2$, $\theta_1 - \mu_A = \mu_A - \theta_2 \implies e(\theta_1, \sigma_q) = e(\theta_2, \sigma_q)$, so $w^\star(\theta_1) = w^\star(\theta_2)$, meaning that $w^\star(\theta)$ and $l^\star(\theta, \mu_A)$ are axisymmetric with respect to $\theta = \mu_A$. Also, $\forall \theta, \mu_A$, $w^\star(\theta)$ and $l^\star(\theta, \mu_A)$ are constant as long as $|\mu_A - \theta|$ is constant. This implies $[l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]$ is centrosymmetric with respect to the point $((\mu_{A1} + \mu_\theta)/2, 0)$. That is, $\forall \theta_1 > \theta_2$, $\theta_1 - (\mu_{A1} + \mu_\theta)/2 = (\mu_{A1} + \mu_\theta)/2 - \theta_2 \implies [l^\star(\theta_1, \mu_{A1}) - l^\star(\theta_1, \mu_\theta)] = -[l^\star(\theta_2, \mu_{A1}) - l^\star(\theta_2, \mu_\theta)] > 0$, which is positive because $l^\star(\theta, \mu_A)$ strictly increases in $|\mu_A - \theta|$ by Proposition 4.

Let $\bar{\mu}$ denote $(\mu_{A1} + \mu_\theta)/2$. Because $\mu_A > \mu_\theta \implies \bar{\mu} > \mu_\theta$, we have $Pr(\theta \leq \bar{\mu}) > Pr(\theta > \bar{\mu})$, and $\forall \theta_1 > \theta_2$, $\theta_1 - \bar{\mu} = \bar{\mu} - \theta_2 \implies \phi\left((\theta_1 - \mu_\theta)/\sigma_\theta\right) < \phi\left((\theta_2 - \mu_\theta)/\sigma_\theta\right)$. Because $[l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]$ is centrosymmetric with respect to the point $(\bar{\mu}, 0)$, these imply $0 < [l^\star(\theta_1, \mu_{A1}) - l^\star(\theta_1, \mu_\theta)]\phi\left((\theta_1 - \mu_\theta)/\sigma_\theta\right) < -[l^\star(\theta_2, \mu_{A1}) - l^\star(\theta_2, \mu_\theta)]\phi((\theta_2 - \mu_\theta)/\sigma_\theta$.

This means that $\forall \theta_1 > \theta_2$, $\theta_1 - \bar{\mu} = \bar{\mu} - \theta_2$, we have

$$[l^\star(\theta_1, \mu_{A1}) - l^\star(\theta_1, \mu_\theta)]\phi\left((\theta_1 - \mu_\theta)/\sigma_\theta\right) + [l^\star(\theta_2, \mu_{A1}) - l^\star(\theta_2, \mu_\theta)]\phi((\theta_2 - \mu_\theta)/\sigma_\theta > 0$$

Hence,

$$\begin{aligned}
&\int_{-\infty}^{\infty} [l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta \\
= &\int_{-\infty}^{\bar{\mu}} [l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta + \int_{\bar{\mu}}^{\infty} [l^\star(\theta, \mu_{A1}) - l^\star(\theta, \mu_\theta)]\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta > 0
\end{aligned}$$

This implies $E[l^\star(\theta, \mu_A)]$ is minimized at $\mu_A = \mu_\theta$.

And because $\dfrac{\partial l^\star(\theta, \mu_A)}{\partial \mu_A}$ is continuous at $\mu_A = \mu_\theta$ and $\sigma_A = \sigma_\theta$, $E[l^\star(\theta, \mu_A)]$ is differentiable at $\mu_A = \mu_\theta$ and $\sigma_A = \sigma_\theta$. Thus, we have $\dfrac{\partial E[l^\star]}{\partial \mu_A}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = 0$.

- By Equation (2.3) and Equation (A.20),

$$l^\star(\theta) = \frac{\sigma_q^{\star 2}(\theta)(\sigma_A^4 + \sigma_q^{\star 2}(\theta)(\mu_A - \theta)^2)}{(\sigma_A^2 + \sigma_q^{\star 2}(\theta))^2} - \frac{\lambda}{2}\ln\left(\frac{\sigma_q^{\star 2}(\theta)}{\sigma_q^{\star 2}(\theta) + \sigma_\theta^2}\right)$$

By the chain rule, $\dfrac{\partial l^\star}{\partial \sigma_A^2} = \dfrac{dl^\star}{d\sigma_q^{\star 2}} \cdot \dfrac{d\sigma_q^{\star 2}}{d\sigma_A^2} + \dfrac{dl^\star}{d\sigma_A^2}$. Because $\sigma_q^{\star 2}$ is optimal, $\dfrac{dl^\star}{d\sigma_q^{\star 2}} = 0$. This implies $\dfrac{\partial l^\star}{\partial \sigma_A^2} = \dfrac{dl^\star}{d\sigma_A^2}$. With some algebra, we can get

$$\frac{dl^\star}{d\sigma_A^2}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = \frac{2\sigma_q^{\star 4}(\sigma_\theta^2 - (\mu_\theta - \theta)^2)}{(\sigma_q^{\star 2} + \sigma_\theta^2)^3}$$

Since $w(\theta) = \sigma_q^2(\theta)/[\sigma_A^2 + \sigma_q^2(\theta)]$, we can rewrite it as

$$\frac{dl^\star}{d\sigma_A^2}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = \frac{dl^\star(\theta)}{d\sigma_A^2}\Big|_{\mu_A = \mu_\theta, \sigma_A = \sigma_\theta} = \frac{2}{\sigma_\theta^2} w^\star(\theta)^2(1 - w^\star(\theta))(\sigma_A^2 - (\mu_\theta - \theta)^2)$$

where $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\Delta}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$ and $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$ by Lemma 6.

And, by definition,

$$E[l^\star]\Big|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} = \int_{-\infty}^{\infty} l^\star(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$$
$$= \int_{|\mu_\theta-\theta|\geq\tau_d} l^\star(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \int_{|\mu_\theta-\theta|<\tau_d} l^\star(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$$

where τ_d is defined in Lemma 6.

When $\mu_A = \mu_\theta$, $l(\theta)$ is symmetric with respect to $\theta = \mu_\theta$, so

$$E[l^\star]\Big|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} = 2\left[\int_{\mu_\theta+\tau_d}^{\infty} l^\star(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \int_{0}^{\mu_\theta+\tau_d} l^\star(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right]$$

And when $w = 1$ we know $l = (\mu_\theta - \theta)^2$, so

$$E[l^\star]\Big|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} = 2\left[\int_{\mu_\theta+\tau_d}^{\infty} l^\star(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \int_{0}^{\mu_\theta+\tau_d} (\mu_\theta-\theta)^2\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right]$$

Thus, by the Leibniz integral rule,

$$\frac{\partial E[l^\star]}{\partial\sigma_A^2}\Big|_{\mu_A=\mu_\theta,\sigma_A=\sigma_\theta} = 2\left[\int_{\mu_\theta+\tau_d}^{\infty} \frac{\partial l^\star(\theta)}{\partial\sigma_A^2}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta - (\mu_\theta-\tau_d)^2\cdot\frac{\partial\tau_d}{\sigma_A^2} + (\mu_\theta-\tau_d)^2\cdot\frac{\partial\tau_d}{\sigma_A^2}\right]$$
$$= 2\left[\int_{\mu_\theta+\tau_d}^{\infty} \frac{\partial l^\star(\theta)}{\partial\sigma_A^2}\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right]$$
$$= 2\left[\int_{\mu_\theta+\tau_d}^{\infty} \frac{2}{\sigma_\theta^2}w^\star(\theta)^2(1-w^\star(\theta))(\sigma_A^2-(\mu_\theta-\theta)^2)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right]$$
$$= \frac{1}{\sigma_\theta^2}\left[\int_{\mu_\theta+\tau_d}^{\infty} w^\star(\theta)^2(1-w^\star(\theta))(\sigma_A^2-(\mu_\theta-\theta)^2)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right]$$

Let $g(\theta) \triangleq w^\star(\theta)^2(1-w^\star(\theta))(\sigma_\theta^2-(\mu_\theta-\theta)^2)$

When $\lambda \geq 2\sigma_\theta^2$, by the proof of Lemma 6 (see the summary at the end of the proof), $\lambda > 2\sigma_\theta^2 \geq \sigma_\theta^2 \implies \tau_d = \sqrt{\sigma_\theta^2/2+\lambda/4} > \sqrt{\sigma_\theta^2/2+2\sigma_\theta^2/4} = \sigma_\theta$. So $g(\theta)$ is always negative for any $\theta > \mu_\theta + \tau_d$. Thus,

$$\int_{\mu_\theta+\tau_d}^{\infty} g(\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta < 0$$

□

Proof. Proof of Theorem 3. Let $w = \sigma_q^2/(\sigma_A^2+\sigma_q^2)$. By Equation (2.1), $\theta_A = (1-w)q+w\mu_A$, where $q = \theta+\epsilon_q$, $\epsilon_q \sim N(0,\sigma_q^2)$ and $\theta \sim N(\mu_\theta,\sigma_\theta^2)$. We further define $w^\star(\theta) = \sigma_q^{\star 2}(\theta)/[\sigma_A^2 + \sigma_q^{\star 2}(\theta)]$. Let $\phi(\cdot)$ denote the probability density function of $N(0,1)$.

$$
\begin{aligned}
E[\theta^\star] &= \int_{|\mu_A-\theta|\le\tau_a}\int_{-\infty}^{\infty}\theta_A^\star\phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right)d\epsilon_q\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\le\tau_a}\int_{-\infty}^{\infty}[(1-w^\star(\theta))q + w^\star(\theta)\mu_A]\phi\left(\frac{\epsilon_q}{\sigma_q^\star(\theta)}\right)d\epsilon_q\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&+ \int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\le\tau_a}[(1-w^\star(\theta))\theta + w^\star(\theta)\mu_A]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\le\tau_a}[(1-w^\star(\theta))(\theta-\mu_A)+\mu_A]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \int_{|\mu_A-\theta|>\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\le\tau_a}[(1-w^\star(\theta))(\theta-\mu_A)+\mu_A]\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \mu_\theta - \int_{|\mu_A-\theta|\le\tau_a}\theta\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\
&= \int_{|\mu_A-\theta|\le\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \mu_\theta
\end{aligned}
$$

This implies that

$$
|E[\theta^\star]-\mu_\theta| = \left|\int_{|\mu_A-\theta|\le\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right| \tag{A.21}
$$

1. (a) First, we want to show

$$
|E[\theta^\star]-\mu_\theta| \le \left|\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right|
$$

Without loss of generality, suppose $\mu_A \ge \mu_\theta$. Then, $Pr(\theta \le \mu_A) \ge Pr(\theta > \mu_A)$, and $\forall\theta_1 > \theta_2$, $\theta_1-\mu_A = \mu_A-\theta_2 \implies \phi((\theta_1-\mu_\theta)/\sigma_\theta) < \phi((\theta_2-\mu_\theta)/\sigma_\theta)$. Because $w^\star(\theta)$ is symmetric with respect to $\theta=\mu_A$, we have $w^\star(\theta_1) = w^\star(\theta_2)$. These imply

$$
0 < -w^\star(\theta_1)(\mu_A-\theta_1)\phi((\theta_1-\mu_\theta)/\sigma_\theta) < w^\star(\theta_2)(\mu_A-\theta_2)\phi((\theta_2-\mu_\theta)/\sigma_\theta)
$$

which means that $\forall\theta_1>\theta_2$, if $\theta_1-\mu_A = \mu_A-\theta_2$, then

$$
w^\star(\theta_2)(\mu_A-\theta_2)\phi((\theta_2-\mu_\theta)/\sigma_\theta) + w^\star(\theta_1)(\mu_A-\theta_1)\phi((\theta_1-\mu_\theta)/\sigma_\theta) > 0
$$

Since $\tau_a > 0$, we can get

$$\int_{\mu_A-\tau_a}^{\mu_A+\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta > 0$$

and

$$\int_{\mu_A+\tau_a}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta + \int_{-\infty}^{\mu_A-\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \geq 0$$

Thus,

$$\begin{aligned}\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta &= \int_{\mu_A-\tau_a}^{\mu_A+\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\ &+ \int_{\mu_A+\tau_a}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\ &+ \int_{-\infty}^{\mu_A-\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\ &> 0\end{aligned}$$

and

$$|E[\theta^\star]-\mu_\theta| = \left|\int_{|\mu_A-\theta|\leq\tau_a} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right| \leq \left|\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right|$$

Let $\lambda_1 > \lambda_2$. By Lemma 13, $\forall\theta$, $w^\star(\theta,\lambda_1) \geq w^\star(\theta,\lambda_2)$. Because $w^\star(\theta)$ is symmetric with respect to $\theta = \mu_A$, $\forall\theta_1 > \theta_2$, $\theta_1 - \mu_A = \mu_A - \theta_2$, then

$$\begin{aligned}&(w^\star(\theta_2,\lambda_1) - w^\star(\theta_2,\lambda_2))(\mu_A-\theta_2)\phi\left((\theta_2-\mu_\theta)/\sigma_\theta\right) \\ \geq\ &-(w^\star(\theta_1,\lambda_1) - w^\star(\theta_1,\lambda_2))(\mu_A-\theta_1)\phi\left((\theta_1-\mu_\theta)/\sigma_\theta\right) \geq 0\end{aligned}$$

This implies

$$\begin{aligned}&\int_{\theta\leq\mu_A} w^\star(\theta,\lambda_1)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta - \int_{\theta\leq\mu_A} w^\star(\theta,\lambda_2)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \\ \geq\ &-\left[\int_{\theta>\mu_A} w^\star(\theta,\lambda_1)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta - \int_{\theta>\mu_A} w^\star(\theta,\lambda_2)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta\right] \geq 0\end{aligned}$$

Rearrange the inequality, we can get

$$\int_{-\infty}^{\infty} w^\star(\theta,\lambda_1)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta \geq \int_{-\infty}^{\infty} w^\star(\theta,\lambda_2)(\mu_A-\theta)\phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right)d\theta$$

Thus,

$$\int_{-\infty}^{\infty} w^{\star}(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

increases in λ.

And because $w^{\star}(\theta, \lambda) \to 1$ as $\lambda \to \infty$, by the monotone convergence theorem (Pugh, 2015), we get the upper bound

$$\int_{-\infty}^{\infty} w^{\star}(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta \leq \mu_A - \mu_\theta$$

Hence, $|E[\theta^{\star}] - \mu_\theta| \leq |\mu_A - \mu_\theta|$.

2. When $\lambda = 0$, for any θ, $w^{\star}(\theta) = 0$, by Equation (A.21), we have $|E[\theta^{\star}] - \mu_\theta| = 0$. And when $\Gamma = 0$, $\tau_a = 0$, $|E[\theta^{\star}] - \mu_\theta| = \left|\int_{|\mu_A - \theta| = 0} w^{\star}(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta\right| = 0$

3. When $\Gamma \to \infty$, by Equation (A.21),

$$|E[\theta^{\star}] - \mu_\theta| = \left|\int_{-\infty}^{\infty} w^{\star}(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta\right|$$

And when $\lambda \to \infty$, $\forall\theta$, $w^{\star}(\theta) \to 1$.

Without loss of generality, suppose $\mu_A \geq \mu_\theta$. In part 1, we have shown that

$$\int_{-\infty}^{\infty} w^{\star}(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

is non-negative and increases in λ. By the monotone convergence theorem (Pugh, 2015), we have

$$\lim_{\lambda \to \infty} \left|\int_{-\infty}^{\infty} w^{\star}(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta\right| = \left|\int_{-\infty}^{\infty} (\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta\right| = |\mu_A - \mu_\theta|$$

Thus, when $\Gamma \to \infty$ and $\lambda \to \infty$, $|E[\theta^{\star}] - \mu_\theta| = |\mu_A - \mu_\theta|$.

4. When $\Gamma \to \infty$, by Equation (A.21),

$$|E[\theta^{\star}] - \mu_\theta| = \left|\int_{-\infty}^{\infty} w^{\star}(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta\right|$$

Without loss of generality, suppose $\mu_A \geq \mu_\theta$. In part 1, we have shown

$$\int_{-\infty}^{\infty} w^\star(\theta)(\mu_A - \theta)\phi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) d\theta$$

is non-negative and increases in λ. Hence, when $\Gamma \to \infty$, $|E[\theta_A^\star] - \mu_\theta|$ increases in λ.

□

A.3 The description of the simulation for the self-training loop.

In this section, we describe the numerical experiment for the self-training loop outlined in Section 2.5. Detailed pseudo code is provided in Algorithm 1, Algorithm 2, Algorithm 3, and Algorithm 4.

Algorithm 1 is the primary algorithm that runs the experiment. There are three key points to highlight: First, for computational tractability, we use a quantization method to discretize all continuous distributions. Specifically, we quantize the population distribution of θ by using the Lloyd-Max algorithm (Gallager et al., 2008), so that we can get a discrete support, $\Theta = \{\theta_1, \ldots, \theta_M\}$ where M is the support size, along with a corresponding probability mass function $\mathbb{P}(\theta),\ \forall\theta \in \Theta$. However, the Lloyd-Max algorithm is not suitable for quantizing the distribution of queries q, because we have to make sure the support of q remains consistent regardless of the mean θ (recall that we define $q = \theta + \epsilon_q$ where $\epsilon_q \sim N(0, \sigma_q^2)$. To address this, we evenly select M_q points from the range $[\underline{\theta} - \Delta_q, \overline{\theta} + \Delta_q]$, where $\underline{\theta}$ and $\overline{\theta}$ are the minimum and maximum values in Θ, respectively. $\Delta_q > 0$ should be large enough to cover most of the support of $N(\theta_A, \sigma_q^2)$ for any $\theta \in \Theta$ and any σ_q that is close to the optimal solution. These points constitute the support of q, denoted by $Q = \{q_1, \ldots, q_{M_q}\}$. The probability mass function is given by $\mathbb{P}(q_i) = \mathbb{P}((q_{i-1} + q_i)/2 < q \geq (q_i + q_{i+1})/2),\ \forall i \in \{2, \ldots, M_q - 1\}$, $\mathbb{P}(q_1) = \mathbb{P}(q \leq (q_1 + q_2)/2)$, and $\mathbb{P}(q_{M_q}) = \mathbb{P}(q > (q_{M_q - 1} + q_{M_q})/2)$ (see Gallager et al. (2008)).

Second, we consider only a finite number of σ_q candidates. In other words, we minimize the utility loss by finding the best σ_q from M_{σ_q} candidates of σ_q rather than by searching

for the optimal σ_q from any non-negative value. This approach maintains computational tractability and stability. Let $\Sigma_q = \{\sigma_1, \ldots, \sigma_{M_{\sigma_q}}\}$ denote the candidate set of q, which should be large enough to yield a solution that is close to the true optimal solution for any $\theta \in \Theta$.

Third, at the end of each iteration, the AI's prior is updated based on the AI outputs. Specifically, the AI's prior is replaced by the distribution of $\theta^\star$: $\pi_A^{t+1}(\theta_i) = \mathbb{P}(\theta^{\star t} = \theta_i),\ \forall \theta_i \in \Theta$. This corresponds to the self-training loop in which the AI learns completely from the AI-generated content in the previous iteration, thereby overriding its prior with the distribution of AI outputs.

Let $\phi(\cdot)$ denote the probability density function of $N(0,1)$. In the base setting, we use $\mu_\theta = 0, \sigma_\theta = 1, M = 1001, T = 100$, where T is the total number of iterations.

Algorithm 2 is used to produce the AI output given the information sent by a user, as depicted in Section 2.3.

Algorithm 3 is used to compute the posterior distribution with respect to the population distribution, π_θ, given q. It helps us to compute the mutual information $e(\theta, \sigma_q)$ in Algorithm 4.

Algorithm 4 is used to compute the utility loss $l(\theta, \sigma_q)$. Note that we compute $I(\theta, \sigma_q)$ by its definition $I(\theta, \sigma_q) = H(\theta) - E_q[H(\theta|q)]$.

Algorithm 1 The steps of the numerical experiment

Input: $\mu_\theta,\ \sigma_\theta,\ T,\ M,\ M_q,\ \Sigma_q,\ \Gamma,\ \lambda$.

Output: $\pi_A^t(\theta_i),\ \forall i \in \{1, 2, \ldots, M\},\ \forall t \in \{1, 2, \ldots, T\}$.

Discretize the population distribution of θ: Apply the Lloyd-Max algorithm to get Θ and $\mathbb{P}(\theta_i),\ \forall \theta_i \in \Theta$.

Discretize the distribution of q: Evenly select M_q points from $[\underline{\theta} - \Delta_q, \overline{\theta} + \Delta_q]$ as Q. Then we compute $\mathbb{P}(q_k|\mu = \theta_i, \sigma = \sigma_j)$ for any $q_k \in Q$, $\theta_i \in \Theta$ and $\sigma_j \in \Sigma_q$.

Initialize the AI's prior: $\pi_A^0(\theta_i) = \pi_\theta(\theta_i),\ \forall \theta_i \in \Theta$

for $t = 0, 2, \ldots, T$ **do**

  for $i = 1, 2, \ldots, M$ **do**

    Find the optimal $\sigma_{q,i}^{\star t} = \arg\min_{\sigma_q \in \Sigma_q} l^t(\theta_i, \sigma_q)$ (Algorithm 4)

    Find the mapping from q_k to θ_A^t: $\theta_A^t(q_k)$ (Algorithm 2)

    Compute the Likelihood: $\mathbb{P}(q_k|\mu = \theta_i, \sigma = \sigma_{q,i}^{\star t}),\ \forall q_k \in Q$

    Compute the conditional distribution of $\theta^{\star t}$ given θ:

    if $l^t(\theta_i, \sigma_{q,i}^{\star t}) > \Gamma$ **then**

      $\mathbb{P}(\theta^{\star t} = \theta_i|\theta = \theta_i) = 1, \mathbb{P}(\theta^{\star t} \neq \theta_i|\theta = \theta_i) = 0.$

    else

      $\mathbb{P}(\theta^{\star t} = \theta_j|\theta = \theta_i) = \sum_{k=1}^{M_q} \mathbb{P}(q_k|\mu = \theta_i, \sigma = \sigma_{q,i}^{\star t}) \mathbf{1}_{\theta_A^t(q_k) = \theta_j},\ \forall \theta_j \in \Theta.$

    end if

  end for

  Compute the distribution of $\theta^{\star t}$ and use it as the new AI prior to the next iteration:

  $\mathbb{P}(\theta^{\star t} = \theta_j) = \sum_{j=1}^{M} \mathbb{P}(\theta^{\star t} = \theta_j|\theta = \theta_i)\mathbb{P}(\theta_i),\ \forall \theta_j \in \Theta$

end for

Algorithm 2 Output θ_A

1: **Input**: π_A, q, σ_q, Θ

2: **Output**: θ_A

3: Compute the likelihood: $\mathbb{P}(q|\mu = \theta, \sigma = \sigma_q),\ \forall \theta \in \Theta$

4: Compute the posterior given q: $\forall \theta \in \Theta, \pi_A(\theta|q, \sigma_q) = \dfrac{\mathbb{P}(q|\mu = \theta, \sigma = \sigma_q)\pi_A(\theta)}{\sum_{\hat{\theta}\in\Theta} \mathbb{P}(q|\mu = \hat{\theta}, \sigma = \sigma_q)\pi_A(\hat{\theta})}$.

5: Compute θ_A minimizing the mean squared error: $\theta_A = \arg\min_{\hat{\theta}\in\Theta} \sum_{\theta\in\Theta} (\hat{\theta} - \theta)^2 \cdot \pi_A(\theta|q, \sigma_q)$

Algorithm 3 Posterior with respect to π_θ

1: **Input**: q, π_θ, σ_q, Θ

2: **Output**: $\pi(\cdot|q, \sigma_q)$

3: Compute the likelihood: $\mathbb{P}(q|\mu = \theta, \sigma = \sigma_q),\ \forall \theta \in \Theta$

4: Compute the posterior given q: $\forall \theta \in \Theta, \pi(\theta|q, \sigma_q) = \dfrac{\mathbb{P}(q|\mu = \theta, \sigma = \sigma_q)\pi_\theta(\theta)}{\sum_{\hat{\theta}\in\Theta} \mathbb{P}(q|\mu = \hat{\theta}, \sigma = \sigma_q)\pi_\theta(\hat{\theta})}$.

Algorithm 4 Compute the utility loss l

1: **Input**: σ_q, θ, π_A, π_θ, S, λ

2: **Output**: l

3: Find the mapping from q to θ_A: $\theta_A(q)$ (Algorithm 2)

4: Compute the likelihood: $\mathbb{P}(q|\mu = \theta, \sigma = \sigma_q),\ \forall \theta \in \Theta$

5: Compute the fidelity error $e(\theta, \sigma_q) = \sum_{q\in Q} [\theta_A(q) - \theta]^2 \mathbb{P}(q|\mu = \theta, \sigma = \sigma_q)$.

6: Compute the mutual information where $\pi(\cdot|q, \sigma_q)$ is given by Algorithm 3

$$I(\theta, \sigma_q) = -\sum_{\theta\in\Theta} \pi_\theta(\theta) \log(\pi_\theta(\theta)) + \sum_{q\in Q} \sum_{\hat{\theta}\in\Theta} \pi(\hat{\theta}|q, \sigma_q) \log(\pi(\hat{\theta}|q, \sigma_q)) \mathbb{P}(q|\mu = \theta, \sigma = \sigma_q)$$

7: Compute $l(\theta, \sigma_q) = e(\theta, \sigma_q) + \lambda I(\theta, \sigma_q)$

A.4 Extensive explanation of Proposition 1: Decomposition of the fidelity error

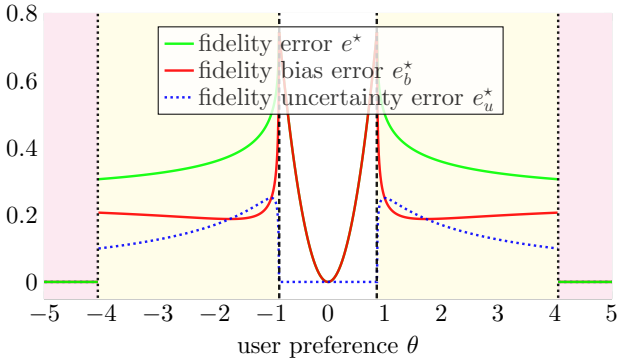


Figure A.2: The black dashed vertical lines are at $d(\theta) = \tau_d$, and the black dotted vertical lines are at $d(\theta) = \tau_a$. The white region indicates the users who simply accept the default output; the yellow region indicates the users interacting with the AI by sending information; the red region indicates the users without using AI. We use $\mu_\theta = 0, \sigma_\theta = 1, \lambda = 1, \Gamma = 1.4$.

To further understand the variation of fidelity error shown in Proposition 1 for the users with $d(\theta) < \tau_a$, we decompose their the fidelity error into a bias and a variance term, $e^\star = Var(\theta^\star|\theta) + [E(\theta^\star|\theta) - \theta]^2$, as introduced in Section 2.3. Again, we call $Var(\theta^\star|\theta)$ the *fidelity uncertainty error* denoted by $e_u^\star$, and $[E(\theta^\star|\theta) - \theta]^2$ the *fidelity bias error* denoted by $e_b^\star$. This decomposition is depicted in Figure A.2. We can see that for the users with $d(\theta) < \tau_d$, the fidelity bias error e_b largely contributes to the fidelity error since they accept the AI's default output without sending any informative signal. At the point of $d(\theta) = \tau_d$, the user starts providing information, leading to a decrease in e_b but an increase in e_u. As the uniqueness further grows, they share more information, resulting in lower fidelity errors. This reduction is primarily driven by the decrease in e_u, since providing more information effectively reduces the noise of the communication but hardly eliminates the inherent difference between the mean and their actual preferences. We formalize this observation in the following Proposition 18.

Proposition 18. For users with $d(\theta) < \tau_a$,

1. The fidelity uncertainty error $e_u^\star$ is zero when $d(\theta) \leq \tau_d$, then increases, and finally decreases in $d(\theta)$.

2. The fidelity bias error $e_b^\star$ increases when $d(\theta) \leq \tau_d$, then decreases and finally increases in $d(\theta)$.

Proof. Proof of Proposition 18. For users with $d(\theta) < \tau_a$, by definition, $e_u(\theta, \sigma_q) = Var(\theta_A|\theta)$ and $e_b(\theta, \sigma_q^2) = [E(\theta_A|\theta) - \theta]^2$. As what we did in the proof of Proposition 17 and Lemma 8, we can show $e_u(\theta, \sigma_q) = w(1-w)\sigma_\theta^2$ and $e_b(\theta, \sigma_q^2) = w^2(\mu_\theta - \theta)^2$, where $w = \dfrac{\sigma_q^2}{\sigma_\theta^2 + \sigma_q^2}$. Thus, $e_u(\theta, \sigma_q^\star(\theta)) = w^\star(\theta)(1 - w^\star(\theta))\sigma_\theta^2$ and $e_b(\theta, \sigma_q^\star(\theta)) = w^{\star 2}(\theta)(\mu_\theta - \theta)^2$, where $w^\star(\theta) = \dfrac{-\sigma_\theta^2 + \sqrt{\sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}}{4((\mu_\theta - \theta)^2 - \sigma_\theta^2)}$ given by Lemma 6.

Then,

1. Fidelity uncertainty error:

   We know that $w^\star(\theta) = 1$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, and $w^\star(\theta) < 1$ for $|\mu_\theta - \theta| \geq \tau_d(\lambda, \sigma_\theta)$. Thus, $e_u(\theta, \sigma_q^\star(\theta)) = 0$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, and $e_u(\theta, \sigma_q^\star(\theta)) > 0$ for $|\mu_\theta - \theta| \geq \tau_d(\lambda, \sigma_\theta)$.

   When $|\mu_\theta - \theta| \geq \tau_d(\lambda, \sigma_\theta)$,

   $$\frac{\partial e_u(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} = \frac{\partial[w^\star(\theta)(1 - w^\star(\theta))\sigma_\theta^2]}{\partial(\mu_\theta - \theta)^2} = \sigma_\theta^2(1 - 2w^\star(\theta))\frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2}$$

   We know $\dfrac{\partial I^\star}{\partial(\mu_\theta - \theta)^2} \geq 0$ by Proposition 1 and $I^\star = -\dfrac{\lambda}{2}\ln w^\star(\theta)$ by Lemma 9. These imply $\dfrac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} \leq 0$. Thus, when $|\mu_\theta - \theta| \geq \tau_d(\lambda, \sigma_\theta)$, the sign of $\dfrac{\partial e_u(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2}$ depends on $(1 - 2w^\star(\theta))$. If $(1 - 2w^\star(\theta)) < 0$ for small $|\mu_\theta - \theta|$, then $e_u(\theta, \sigma_q^\star(\theta))$ first increases and then decreases in $|\mu_\theta - \theta|$; if $(1 - 2w^\star(\theta)) \geq 0$ for any $|\mu_\theta - \theta|$, monotonically decreases in $|\mu_\theta - \theta|$. (Notice that $(1 - 2w^\star(\theta))$ is always positive for sufficiently large $|\mu_\theta - \theta|$, because $w^\star(\theta) \to 0$ as $|\mu_\theta - \theta| \to \infty$.)

Hence, we either have $e_u(\theta, \sigma_q^\star(\theta)) = 0$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, first increases and then decreases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| \geq \tau_d(\lambda, \sigma_\theta)$; or $e_u(\theta, \sigma_q^\star(\theta)) = 0$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, then there is a jump at $|\mu_\theta - \theta| = \tau_d(\lambda, \sigma_\theta)$ ($e_u(\theta, \sigma_q^\star(\theta))$ jumps to a positive value), and then $e_u(\theta, \sigma_q^\star(\theta))$ monotonically decreases in $|\mu_\theta - \theta|$.

2. Fidelity bias error:

   We know that $w^\star(\theta) = 1$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, so $e_b(\theta, \sigma_q^\star(\theta)) = (\mu_\theta - \theta)^2$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, which is increasing in $|\mu_\theta - \theta|$.

   At $|\mu_\theta - \theta| = \tau_d(\lambda, \sigma_\theta)$, since $w^\star(\theta) < 1$ is optimal, we have $e_b(\theta, \sigma_q^\star(\theta)) = e(\theta, \sigma_q^{\star 2}(\theta)) - e_u(\theta, \sigma_q^\star(\theta)) < e(\theta, \sigma_q^{\star 2}(\theta)) < (\mu_\theta - \theta)^2$. Thus, $e_b(\theta, \sigma_q^\star(\theta))$ decreases at $|\mu_\theta - \theta| = \tau_d(\lambda, \sigma_\theta)$.

   When $|\mu_\theta - \theta| > \tau_d(\lambda, \sigma_\theta)$,

$$\begin{aligned}
\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial(\mu_\theta - \theta)^2} &= \frac{\partial[w^{\star 2}(\theta)(\mu_\theta - \theta)^2]}{\partial(\mu_\theta - \theta)^2} \\
&= 2(\mu_\theta - \theta)^2 w^\star(\theta) \frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} + w^{\star 2}(\theta) \\
&= w^\star(\theta) \left[2(\mu_\theta - \theta)^2 \frac{\partial w^\star(\theta)}{\partial(\mu_\theta - \theta)^2} + w^\star(\theta) \right]
\end{aligned}$$

   Substitute Equation (A.4) into the above equation

$$= w^\star(\theta) \left[2(\mu_\theta - \theta)^2 \cdot \frac{\sigma_\theta^2 \sqrt{\Delta} - \sigma_\theta^4 - 2\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2} + w^\star(\theta) \right]$$

   where $\Delta = \sigma_\theta^4 + 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)$

   With some simplifications

$$= w^\star(\theta) \cdot \frac{\sigma_\theta^2[((\mu_\theta - \theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda((\mu_\theta - \theta)^2 - \sigma_\theta^2)]}{4\sqrt{\Delta}((\mu_\theta - \theta)^2 - \sigma_\theta^2)^2}$$

   Let $d(\theta) = |\mu_\theta - \theta|$. Then,

$$\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2} = w^\star(\theta) \cdot \frac{\sigma_\theta^2[(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda(d(\theta)^2 - \sigma_\theta^2)]}{4\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2}$$

Now, we want to show that when $d(\theta) \geq \tau_d(\lambda, \sigma_\theta)$ and $d(\theta)$ is finite, $\dfrac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ has at most one zero point with respect to $d(\theta)$. This is equivalent to showing

$$\frac{(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda(d(\theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2}$$

has at most one zero point with respect to $d(\theta)$. Let $\widehat{d(\theta)}$ denote a solution of $d(\theta)$ such that

$$\frac{(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda(d(\theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2} = 0$$

First, let the nominator $(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda(d(\theta)^2 - \sigma_\theta^2) = 0$, we get:

$$\begin{aligned}
& (d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) = 4\lambda(d(\theta)^2 - \sigma_\theta^2) \\
\implies & (d(\theta)^2 + \sigma_\theta^2)\sqrt{\Delta} = 4\lambda(d(\theta)^2 - \sigma_\theta^2) + \sigma_\theta^2(d(\theta)^2 + \sigma_\theta^2) \\
\implies & \sqrt{\Delta} = \frac{4\lambda(d(\theta)^2 - \sigma_\theta^2) + \sigma_\theta^2(d(\theta)^2 + \sigma_\theta^2)}{d(\theta)^2 + \sigma_\theta^2} \\
\implies & 1 = \frac{4\lambda(d(\theta)^2 - \sigma_\theta^2)}{(d(\theta)^2 + \sigma_\theta^2)^2} + \frac{2\sigma_\theta^2}{d(\theta)^2 + \sigma_\theta^2} \\
\implies & (d(\theta)^2 + \sigma_\theta^2)^2 = 4\lambda(d(\theta)^2 - \sigma_\theta^2) + 2\sigma_\theta^2(d(\theta)^2 + \sigma_\theta^2) \\
\implies & (d(\theta)^2 - \sigma_\theta^2)(d(\theta)^2 + \sigma_\theta^2 - 4\lambda) = 0
\end{aligned}$$

So the candidates of $\widehat{d(\theta)}$ are $\widehat{d(\theta)} = \sigma_\theta$, and $\widehat{d(\theta)} = \sqrt{4\lambda - \sigma_\theta^2}$ if $4\lambda \geq \sigma_\theta^2$

Furthermore, by using L'Hôpital's rule, one can get

$$\lim_{\widehat{d(\theta)} \to \sigma_\theta} \frac{(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda(d(\theta)^2 - \sigma_\theta^2)}{\sqrt{\Delta}(d(\theta)^2 - \sigma_\theta^2)^2} = \frac{2\sigma_\theta^2\lambda - 4\lambda^2}{\sigma_\theta^6}$$

which is zero if and only if $\sigma_\theta^2 = 2\lambda$. This means when $\sigma_\theta^2 \neq 2\lambda$, there is no real $\widehat{d(\theta)}$ if $4\lambda < \sigma_\theta^2$ or $\widehat{d(\theta)} = \sqrt{4\lambda - \sigma_\theta^2}$ if $4\lambda \geq \sigma_\theta^2$.

And when $\sigma_\theta^2 = 2\lambda$, $\widehat{d(\theta)} = \sqrt{4\lambda - \sigma_\theta^2} = \sigma_\theta$, so we also only have one solution for $\widehat{d(\theta)}$. Thus, $\dfrac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ has at most one zero point with respect to $d(\theta)$.

In addition, if $d(\theta) > \sigma_\theta$, $(d(\theta)^2 + \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda(d(\theta)^2 - \sigma_\theta^2) > (d(\theta)^2 - \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2) - 4\lambda(d(\theta)^2 - \sigma_\theta^2) = (d(\theta)^2 - \sigma_\theta^2)(\sqrt{\Delta} - \sigma_\theta^2 - 4\lambda)$, which is positive if $d(\theta) >$

$\max\{\sigma_\theta, \sqrt{|4\lambda - \sigma_\theta^2|}\}$. This means that for any $d(\theta) > \max\{\sigma_\theta, \sqrt{|4\lambda - \sigma_\theta^2|}\}$, $\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ is positive. Because we have shown $\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ has at most one zero point with respect to $d(\theta)$, the intermediate value theorem implies that when $|\mu_\theta - \theta| > \tau_d(\lambda, \sigma_\theta)$, $\frac{\partial e_b(\theta, \sigma_q^\star(\theta))}{\partial d(\theta)^2}$ is either always positive, or negative for small $|\mu_\theta - \theta|$ and then positive for large $|\mu_\theta - \theta|$.

Hence, we either have $e_b(\theta, \sigma_q^\star(\theta))$ first increases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, then decreases and finally increases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| \geq \tau_d(\lambda, \sigma_\theta)$; or $e_b(\theta, \sigma_q^\star(\theta))$ first increases in $|\mu_\theta - \theta|$ for $|\mu_\theta - \theta| < \tau_d(\lambda, \sigma_\theta)$, then there is a jump at $|\mu_\theta - \theta| = \tau_d(\lambda, \sigma_\theta)$ ($e_b(\theta, \sigma_q^\star(\theta))$ jumps to a smaller value), and then $e_b(\theta, \sigma_q^\star(\theta))$ monotonically increases in $|\mu_\theta - \theta|$.

□

A.5 More Detailed Version of Theorem 1.

In this section, we present a more detailed description of Theorem 1 :

Theorem 12 (Full version of Theorem 1)**.** *When $\Gamma \to +\infty$, the variance of the population output is lower than the variance of the population preferences, $Var(\theta^\star) < Var(\theta)$, and strictly decreases in the cost of human-AI interactions λ. When $\Gamma < +\infty$, $\lim_{\lambda \to 0} Var(\theta^\star) = Var(\theta)$ and $\lim_{\lambda \to +\infty} Var(\theta^\star) < Var(\theta)$. In addition, $Var(\theta^\star) < Var(\theta)$ if $\lambda \geq \sigma_q^2/2$ or $\Gamma \leq \hat{\Gamma}$ or $\Gamma \geq \tilde{\Gamma}$ for some $\hat{\Gamma} > 0, \tilde{\Gamma} > 0$.*

The full proof is provided in Appendix A.2.1. The last sentence in this detailed version is the additional part compared with the version that we presented in the main text. In particular, we show that the population variance of the output is strictly less than the population variance of the preferences if λ is sufficiently large or Γ is outside an interval $(\hat{\Gamma}, \tilde{\Gamma})$.

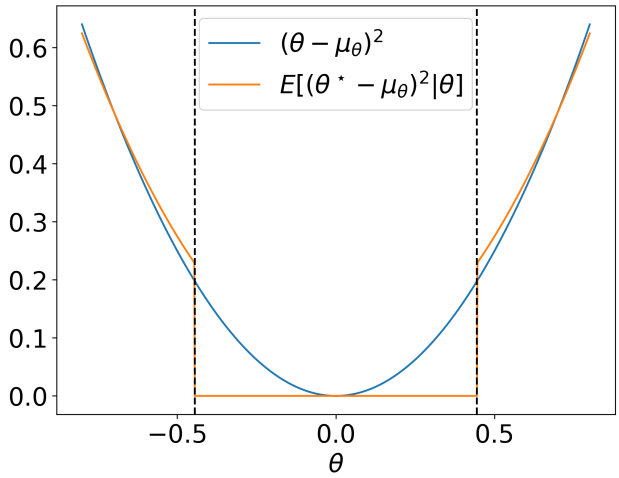


Figure A.3: We use $\mu_\theta = 0, \sigma_\theta = 1, \lambda = 0.1$. The black dashed vertical lines are at $d(\theta) = \tau_d$. The orange curve is above the blue curve for some θ with $d(\theta) > \tau_d$ but close to τ_d, showing that $E[(\theta^\star - \mu_\theta)^2|\theta] \geq (\theta - \mu_\theta)^2$ for these user θ.

In Figure A.3, we show why it is possible that the population variance of the output can be larger than the population variance of the preferences when $\lambda < \sigma_q^2/2$ and $\Gamma \in (\hat{\Gamma}, \tilde{\Gamma})$. By the tower property of conditional expectation, we know $Var(\theta^\star) = E[E[(\theta^\star - \mu_\theta)^2|\theta]]$ (Notice that $E[\theta^\star] = \mu_\theta$ is shown in the proof of Theorem 1). So if $E[(\theta^\star - \mu_\theta)^2|\theta] < (\theta - \mu_\theta)^2$, we must have $Var(\theta^\star) < Var(\theta)$. However, it is possible that $E[(\theta^\star - \mu_\theta)^2|\theta] \geq (\theta - \mu_\theta)^2$ for some θ whose $d(\theta)$ are close to τ_d. Since τ_d is the root of a transcendental equation, it is complicated to find the closed form of this region. Despite this possibility, it is actually hard to find a scenario such that $Var(\theta^\star) > Var(\theta)$ in our numerical tests.

Intuitively, the users with $d(\theta) > \tau_d$ but close to τ_d will send a small amount of information. Since the information is always noisy, it will also add more randomness and uncertainty to the outputs. So these users have a higher $E[(\theta^\star - \mu_\theta)^2|\theta]$ and can "contribute" more to the population variance of outputs. However, there are also many users who simply accept the default outputs (i.e., $d(\theta) < \tau_d$) and users with preferences that are far from the mean. They have a lower $E[(\theta^\star - \mu_\theta)^2|\theta]$, thereby reducing the population variance of outputs. These two counter-forces interact with each other, leading to a change in the variance.

APPENDIX B

Autonomous Vehicles in Ride-Hailing and the Threat of Spatial Inequalities

B.1 Proof of the Main Results

B.1.1 Proof of Lemma 2.

Proof. Proof of Lemma 2. First, we want to show that given N_A and N_H, for any idling policy, there is a non-idling policy that can produce a higher profit. To prove this, let's suppose some AVs may be idled (the case of idling HVs is exactly the same). Let $N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$ be the average number of AVs in a system with two types of vehicles under a policy π which may idle AVs at some location j, where the arrival rate of vehicles is $\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H$. We call π as an idling AV policy.

Let $\pi^\star$ denote a non-idling AV policy. We argue that $N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) > N_A^{\pi^\star}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$ for any idling policy π. Indeed, given arrival rates to the system, $\boldsymbol{\lambda}_A$, $\boldsymbol{\lambda}_H$, when AVs are available, $\pi^\star$ never rejects a request while π might reject requests at location j. The rejection in π increases the queue size. This implies that the average number of AVs in the system under π is larger than under $\pi^\star$. Moreover, for any idling policy π, $N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$ increases with $\lambda_{A,j}$ of $\boldsymbol{\lambda}_A$.

Given the constraints $N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) = N_A$, there must exist $\lambda^\star_{A,j} > \lambda_{A,j}, \lambda^\star_{H,j} \geqslant \lambda_{H,j}$ such that $N_A^{\pi^\star}(\boldsymbol{\lambda}^\star_A, \boldsymbol{\lambda}^\star_H) = N_A$, because $N_A^{\pi^\star}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) < N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) = N_A$ and $N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$ increase with $\lambda_{A,j}$. Therefore, we can find a non-idling policy $\pi^\star$ such that $N_A^{\pi^\star}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) =$

$N_A^\pi(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) = N_A$ and producing a higher profit. That is,

$$\begin{aligned}&(P_A - c_A)\sum_{j=1}^{L}\tau_j\lambda_{A,j}^\star + (1-\gamma)P_H\sum_{j=1}^{L}\tau_j\lambda_{H,j}^\star\\ \geqslant&(P_A - c_A)\sum_{j=1}^{L}\tau_j\lambda_{A,j} + (1-\gamma)P_H\sum_{j=1}^{L}\tau_j\lambda_{H,j}\end{aligned}$$

The analysis is the same for the case of idling HVs.

The above analysis means that given N_A and N_H, for any idling policy π which satisfies the wage equilibrium with $\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H$, there exists a non-idling policy $\pi^\star$ which has a higher profit than π with $\boldsymbol{\lambda}_A^\star, \boldsymbol{\lambda}_H^\star$.

Second, assume π is feasible in Problem $(\mathcal{M})$ so that the wage equilibrium is satisfied (i.e. $\gamma\sum_{j=1}^{L}P_H\tau_j\lambda_{H,j} = rN_H$, where N_H is the average number of HVs under π.), then we want to show that there exists an equilibrium number of HVs, $N_H^\star$, under the above non-idling policy $\pi^\star$, such that $N_H^\star \geqslant N_H$. First notice that if $\lambda_{H,j} = \lambda_{H,j}^\star$ for all j, the wage equilibrium must be satisfied, since π is feasible in Problem $(\mathcal{M})$. That is, $N_H^\star = N_H$ if $\lambda_{H,j} = \lambda_{H,j}^\star$ for all j.

If $\lambda_{H,j} < \lambda_{H,j}^\star$ at some location j, then we have:

$$\gamma P_H\sum_{j=1}^{L}\tau_j\lambda_{H,j}^\star > \gamma P_H\sum_{j=1}^{L}\tau_j\lambda_{H,j} = rN_H \tag{B.1}$$

This means that more HVs will enter the market. And by Little's law:

$$\sum_{j=1}^{L}(\tau_j + W_j(\lambda_{A,j}, \lambda_{H,j}))(\lambda_{A,j}^\star + \lambda_{H,j}^\star) = N_A + N_H$$

We know that for any non-idling policy, the expected waiting function at location j is $W_j(\lambda_{A,j}, \lambda_{H,j}) = 1/(\mu_j - \lambda_{A,j} - \lambda_{H,j})$, so the above equation becomes: $\sum_{j=1}^{L}(\tau_j + 1/(\mu_j - \lambda_{A,j}^\star - \lambda_{H,j}^\star))(\lambda_{A,j}^\star + \lambda_{H,j}^\star) = N_A + N_H$. Substitute Equation (B.1) with N_H in the above equation, we have $\sum_{j=1}^{L}(\tau_j + 1/(\mu_j - \lambda_{A,j}^\star - \lambda_{H,j}^\star))(\lambda_{A,j}^\star + \lambda_{H,j}^\star) - (\gamma P_H/r)\sum_{j=1}^{L}\tau_j\lambda_{H,j}^\star < N_A$. The left-hand side of the above inequality is continuous and converges to ∞ as we increase

$\lambda^\star_{H,j}$ up to $\mu_j - \lambda^\star_{A,j}$, so we must be able to find $\tilde{\lambda}_{H,j} > \lambda^\star_{H,j}$ and $N^\star_H > N_H$ such that $\sum_{j=1}^L (\tau_j + 1/(\mu_j - \lambda^\star_{A,j} - \tilde{\lambda}_{H,j}))(\lambda^\star_{A,j} + \tilde{\lambda}_{H,j}) - (\gamma P_H/r)\sum_{j=1}^L \tau_j \tilde{\lambda}_{H,j} = N_A$, and $\gamma \sum_{j=1}^L P_H \tau_j \tilde{\lambda}_{H,j} = r N^\star_H$. In this case, both Little's law and the wage equilibrium are satisfied, and the profit under $\pi^\star$ is higher than that under π.

□

B.1.2 Proof of Proposition 6 and Proposition 7.

By proving Proposition 6, we will show that any optimal solution to Problem ($\mathcal{M}$) is feasible in Problem ($\mathcal{M}'$). Then, to prove Proposition 7, we also need to demonstrate any optimal solution to Problem ($\mathcal{M}'$) is feasible in Problem ($\mathcal{M}$). That is, we will first show that the arrival rates implied by any optimal solution to Problem ($\mathcal{M}$) are bounded by the arrival rates under the full prioritization policies. Then, we will show that the arrival rates implied by any optimal solution to Problem ($\mathcal{M}'$) can be achieved by an implementable policy π in Problem ($\mathcal{M}$). To this end, let us begin with some auxiliary reformulations and definitions.

According to Lemma 2, Problem ($\mathcal{M}$) can be rewritten as

$$
\begin{aligned}
\sup_{\pi, N_H, \lambda_{i,j} \geq 0, \lambda_{A,j} + \lambda_{H,j} < \mu_j} \quad & \sum_{j=1}^L (P_A - c_A)\tau_j \cdot \lambda_{A,j} + (1-\gamma)\sum_{j=1}^L P_H \tau_j \cdot \lambda_{H,j} \\
\text{s.t.} \quad & \sum_{j=1}^L (\tau_j + W_{i,j}(\pi, \lambda_{A,j}, \lambda_{H,j}))\lambda_{i,j} = N_i, \qquad i \in \{A, H\} \\
& \gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j} = r N_H \\
& \sum_{j=1}^L (\tau_j + \frac{1}{\mu_j - (\lambda_{A,j} + \lambda_{H,j})})(\lambda_{A,j} + \lambda_{H,j}) = N_H + N_A
\end{aligned}
$$

The extra constraint indicates any optimal solution must be non-idling.

Let $\lambda_j = \lambda_{A,j} + \lambda_{H,j},\ j \in \{1, \ldots, L\}$, and replace $\lambda_{H,j}$ by $\lambda_j - \lambda_{A,j}$:

$$\sup_{\pi, N_H \geqslant 0, \mu_j > \lambda_j \geqslant \lambda_{A,j} \geq 0} \quad \sum_{j=1}^{L} (P_A - c_A - (1-\gamma)P_H)\tau_j \cdot \lambda_{A,j} + (1-\gamma)\sum_{j=1}^{L} P_H \tau_j \cdot \lambda_j \tag{B.2}$$

$$\text{s.t.} \quad \sum_{j=1}^{L} (\tau_j + W^{\pi}_{A,j}(\lambda_{A,j}, \lambda_j))\lambda_{A,j} = N_A \tag{B.2a}$$

$$\sum_{j=1}^{L} (\tau_j + W^{\pi}_{H,j}(\lambda_{A,j}, \lambda_j))(\lambda_j - \lambda_{A,j}) = N_H \tag{B.2b}$$

$$\gamma P_H \sum_{j=1}^{L} \tau_j (\lambda_j - \lambda_{A,j}) = r N_H \tag{B.2c}$$

$$\sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j})\lambda_j = N_H + N_A \tag{B.2d}$$

It is unclear how to solve the above problem due to the existence of the unknown waiting time functions. However, we can use the achievable region approach to transform the original problem into a problem in terms of the performance metric that we choose. That is, we want to show that the above Problem (B.2) is equivalent to the following Problem (B.3), and then show that Problem ($\mathcal{M}'$) is equivalent to Problem (B.3).

$$\max_{\mu_j > \lambda_j \geqslant \lambda_{A,j} \geqslant 0, N_H \geqslant 0} \quad \sum_{j=1}^{L} (P_A - c_A - (1-\gamma)P_H)\tau_j \cdot \lambda_{A,j} + (1-\gamma)\sum_{j=1}^{L} P_H \tau_j \cdot \lambda_j \tag{B.3}$$

$$\text{s.t.} \quad \gamma P_H \sum_{j=1}^{L} \tau_j (\lambda_j - \lambda_{A,j}) = r N_H \tag{B.3a}$$

$$\sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j})\lambda_j = N_H + N_A \tag{B.3b}$$

$$\sum_{j=1}^{L} \tau_j \lambda_j \geqslant \sum_{j=1}^{L} \tau_j \lambda_j^{\dagger} \tag{B.3c}$$

$$\sum_{j=1}^{L} \tau_j \lambda_j \leqslant \sum_{j=1}^{L} \tau_j \lambda_j^{\ddagger} \tag{B.3d}$$

where $\{\lambda_j^{\dagger}\}_{j=1}^{L}$ is the optimal arrival rates when AVs are fully prioritized, and $\{\lambda_j^{\ddagger}\}_{j=1}^{L}$ is the optimal arrival rates when HVs are fully prioritized. Specifically, they are derived by the

following steps:

1. Fully prioritize AVs

   (a) By fully prioritizing AVs, AVs are not affected by HVs at all, and are distributed with the objective of maximizing $\sum_{j=1}^{L} \tau_j \lambda_{A,j}$. Thus, the waiting time of AVs is $1/(\mu_j - \lambda_{A,j})$ at each location j. To find the optimal solution to AVs, we solve the following optimization problem:

$$\begin{aligned} \max_{\lambda_{A,j} \in [0,\mu_j)} \quad & g(\boldsymbol{\lambda}_A) = \sum_{j=1}^{L} \tau_j \lambda_{A,j} \\ \text{s.t.} \quad & \sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_{A,j}}) \lambda_{A,j} = N_A \end{aligned} \tag{B.4}$$

   Let $\boldsymbol{\lambda}_A^\dagger = \{\lambda_{A,j}^\dagger\}_{j=1}^L$ denote the optimal solution to the above problem.

   (b) Given $\boldsymbol{\lambda}_A^\dagger$, we use the remaining capacity to dispatch HVs. In order to maximize the profit, we should maximize the arrival rate of HVs, since the arrival rate of AVs is fixed now. Given the fixed arrival rate of AVs, maximizing the arrival rate of HVs is equivalent to maximizing the overall arrival rates. That is, to find the optimal policy for HVs, we maximize the arrival rates of HVs, which is equivalent to maximizing the overall arrival rates given $\boldsymbol{\lambda}_A^\dagger$:

$$\begin{aligned} \max_{\lambda_j \in [0,\mu_j), N_H \geqslant 0} \quad & z(\boldsymbol{\lambda}) = \sum_{j=1}^{L} \tau_j \lambda_j \\ \text{s.t.} \quad & \sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j}) \lambda_j = N_A + N_H \\ & r N_H = \gamma P_H \sum_{j=1}^{L} \tau_j (\lambda_j - \lambda_{A,j}^\dagger) \end{aligned} \tag{B.5}$$

   Let $\boldsymbol{\lambda}^\dagger = \{\lambda_j^\dagger\}_{j=1}^L$, $N_H^\dagger = \gamma P_H \sum_{j=1}^{L} \tau_j (\lambda_j^\dagger - \lambda_{A,j}^\dagger)/r$ denote the optimal solution to Problem (B.5).

2. Fully prioritize HVs

(a) By fully prioritizing HVs, HVs are not affected by AVs at all, and are distributed with the objective of maximizing $\sum_{j=1}^{L} \tau_j \lambda_{H,j}$. Thus, the waiting time of HVs is $\frac{1}{\mu_j - \lambda_{H,j}}$ at each location j. And to find the optimal allocation of HVs, we need to solve the following maximization problem:

$$\begin{aligned} \max_{\lambda_{H,j} \in [0,\mu_j), N_H \geqslant 0} \quad & g(\boldsymbol{\lambda}_H) = \sum_{j=1}^{L} \tau_j \lambda_{H,j} \\ \text{s.t.} \quad & \sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_{H,j}}) \lambda_{H,j} = N_H \\ & r N_H = \gamma P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j} \end{aligned} \tag{B.6}$$

Let $\boldsymbol{\lambda}_H^{\ddagger} = \{\lambda_{H,j}^{\ddagger}\}_{j=1}^{L}$ to denote the optimal solution to HVs derived from the above problem, and let $N_H^{\ddagger} = \gamma P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j}^{\ddagger} / r$ to denote the optimal number of HVs that is implied by the wage equilibrium.

(b) Given $\boldsymbol{\lambda}_H^{\ddagger}$, we use the remaining capacity to dispatch AVs. Again, to find the optimal allocation for AVs, we maximize the arrival rate of AVs, which is equivalent to maximizing the overall arrival rates given the arrival rates of HVs $\boldsymbol{\lambda}_H^{\ddagger}$:

$$\begin{aligned} \max_{\lambda_j \in [0,\mu_j)} \quad & z(\boldsymbol{\lambda}) = \sum_{j=1}^{L} \tau_j \lambda_j \\ \text{s.t.} \quad & \sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j}) \lambda_j = N_A + N_H^{\ddagger} \end{aligned} \tag{B.7}$$

Let $\boldsymbol{\lambda}^{\ddagger} = \{\lambda_j^{\ddagger}\}_{j=1}^{L}$ denote the optimal solution to Problem (B.7). Note that it is not necessary to impose $\lambda_j \geqslant \lambda_{H,j}^{\ddagger}$ because $\boldsymbol{\lambda}^{\ddagger} = \boldsymbol{\lambda}_H^{\ddagger}$ is a feasible solution to Problem (B.7) with $N_A = 0$, and the optimal $\lambda_j^{\ddagger}$ must increase in N_A as shown in the proof of Lemma 15.

The following steps require some auxiliary lemmas to complete. Please refer to Appendix B.2.1 and Appendix B.2.2 for the details.

Now, we are ready to complete the proof of Proposition 7. First, we will show that any optimal solution in Problem (B.2) is feasible in Problem (B.3), which implies Proposition 6, and then show that Problem (B.3) is equivalent to Problem ($\mathcal{M}'$). And finally, we will show that any optimal solution to Problem ($\mathcal{M}'$) is feasible in Problem (B.2). Since Problem (B.2) is another formulation of Problem ($\mathcal{M}$), these steps will prove that Problem ($\mathcal{M}$) is equivalent to Problem ($\mathcal{M}'$).

Proof. Proof of Proposition 6. Now let us show that any optimal solution in Problem (B.2) is feasible in Problem (B.3).

First, we want to show any optimal solution in Problem (B.2) is feasible in Problem (B.3). It is sufficient to show any optimal solution in Problem (B.2) satisfies the inequality constraints (B.3c) and (B.3d). Let $(\boldsymbol{\lambda}^\star = \{\lambda_j^\star\}_{j=1}^L,\ \boldsymbol{\lambda}_A^\star = \{\lambda_{A,j}^\star\}_{j=1}^L,\ N_H^\star)$ denote an optimal solution to Problem (B.2).

1. $\sum_{j=1}^L \tau_j \lambda_j^\star \geqslant \sum_{j=1}^L \tau_j \lambda_j^\dagger$:

   Since $(\boldsymbol{\lambda}^\star = \{\lambda_j^\star\}_{j=1}^L,\ \boldsymbol{\lambda}_A^\star = \{\lambda_{A,j}^\star\}_{j=1}^L,\ N_H^\star)$ is an optimal solution without the constraint of fully prioritizing AVs, its objective value must be not less than that of any other solutions:

$$(1-\gamma)P_H \sum_{j=1}^L \tau_j \lambda_j^\star + (P_A - c_A - (1-\gamma)P_H) \sum_{j=1}^L \tau_j \lambda_{A,j}^\star$$

$$\geqslant (1-\gamma)P_H \sum_{j=1}^L \tau_j \lambda_j^\dagger + (P_A - c_A - (1-\gamma)P_H) \sum_{j=1}^L \tau_j \lambda_{A,j}^\dagger$$

$$\implies (1-\gamma)P_H \sum_{j=1}^L \tau_j (\lambda_j^\star - \lambda_j^\dagger) \geqslant (P_A - c_A - (1-\gamma)P_H) \sum_{j=1}^L \tau_j (\lambda_{A,j}^\dagger - \lambda_{A,j}^\star)$$

   Because of *Lemma* 16, the maximum arrival rate of AVs can be achieved only if we fully prioritize AVs. And since $\boldsymbol{\lambda}_A^\dagger$ is the optimal arrival rate of AVs when we fully prioritize AVs, $(P_A - c_A)\sum_{j=1}^L \tau_j \lambda_{A,j}^\dagger$ is the maximum revenue of AVs. Thus, $\sum_{j=1}^L \tau_j (\lambda_{A,j}^\dagger - \lambda_{A,j}^\star) \geqslant 0$. This implies $\sum_{j=1}^L \tau_j (\lambda_j^\star - \lambda_j^\dagger) \geqslant 0$.

2. $\sum_{j=1}^{L} \tau_j \lambda_j^\star \leqslant \sum_{j=1}^{L} \tau_j \lambda_j^\ddagger$:

   First, because of *Lemma* 16, $\sum_{j=1}^{L} \tau_j \lambda_{H,j}^\ddagger$ is the maximum arrival rate of HVs that can be achieved in equilibrium and feasible in Problem $(\mathcal{M})$; and its implied number of HVs $N_H^\ddagger = \gamma P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j}^\ddagger / r$ is the maximum number of HVs that can be achieved at equilibrium and feasible in Problem $(\mathcal{M})$.

   In addition, note that Problem (B.7) is equivalent to Problem (B.11) with $N = N_A + N_H^\ddagger$, and any optimal arrival rates $\boldsymbol{\lambda}^\star = \{\lambda_j^\star\}_{j=1}^{L}$ in Problem (B.2) are feasible in Problem (B.11) with $N = N_A + N_H^\star$. Since $N_H^\ddagger$ is the maximum number of HVs that can be achieved in equilibrium, we have $N_H^\ddagger \geqslant N_H^\star$ and $h(N_A + N_H^\ddagger) \geqslant h(N_A + N_H^\star)$ by Lemma 15. Therefore, we must have $\sum_{j=1}^{L} \tau_j \lambda_j^\star \leqslant \sum_{j=1}^{L} \tau_j \lambda_j^\ddagger$.

Hence, any optimal solution in Problem (B.2) is feasible in Problem (B.3). Since Problem (B.2) is another formulation of Problem $(\mathcal{M})$, this means any optimal solution in Problem $(\mathcal{M})$ is feasible in Problem (B.3), which implies Proposition 6.

□

Proof. Proof of Proposition 7.

To complete the proof of Proposition 7, we continue to show the equivalence between Problem (B.3) and Problem $(\mathcal{M}')$, and then show that any optimal solution to Problem $(\mathcal{M}')$ is feasible in Problem (B.2).

Equivalence between Problem (B.3) **and Problem** $(\mathcal{M}')$**.** Notice that since $\lambda_{A,j}, \lambda_j$ and N_H in Problem (B.3) are all decision variables, we can combine Constraint (B.3a) and (B.3b):

$$N_A + \frac{\gamma P_H \sum_{j=1}^{L} \tau_j (\lambda_j - \lambda_{A,j})}{r} = \sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j}) \lambda_j$$

which implies

$$\sum_{j=1}^{L} \tau_j \lambda_{A,j} = \frac{rN_A + (\gamma P_H - r)\sum_{j=1}^{L} \tau_j \lambda_j - r\sum_{j=1}^{L} \frac{\lambda_j}{\mu_j - \lambda_j}}{\gamma P_H} \tag{B.8}$$

Replace $\sum_{j=1}^{L} \tau_j \lambda_{A,j}$ with the above equation, Problem (B.3) is transformed into Problem $(\mathcal{M}')$:

$$\begin{aligned} \max_{\pi, \lambda_j \in [0,\mu_j)} & \sum_{j=1}^{L} (p - \hat{r})\tau_j \lambda_j + \hat{r} N_A - \hat{r} \sum_{j=1}^{L} \frac{\lambda_j}{\mu_j - \lambda_j} \\ \text{s.t.} \quad & \sum_{j=1}^{L} \tau_j \lambda_j \geqslant \sum_{j=1}^{L} \tau_j \lambda_j^{\dagger} \\ & \sum_{j=1}^{L} \tau_j \lambda_j \leqslant \sum_{j=1}^{L} \tau_j \lambda_j^{\ddagger} \end{aligned} \tag{$\mathcal{M}'$}$$

where $p \triangleq P_A - c_A, \hat{r} \triangleq r[P_A - c_A - (1-\gamma)P_H]/(\gamma P_H)$. Note that it is unnecessary to add a constraint of $\lambda_j \geqslant \lambda_{A,j}$, because we will show that for any optimal solution to Problem $(\mathcal{M}')$, there exists an implementable policy in Problem (B.2).

The above equivalence implies that any optimal solution to Problem (B.2) is feasible in Problem $(\mathcal{M}')$. To complete this proof, we also need to show that any optimal solution to Problem $(\mathcal{M}')$ is feasible in Problem (B.2). In fact, since Problem $(\mathcal{M}')$ is a convex problem with a strictly concave objective, it must have a unique optimal solution. With a slight abuse of notation, we use $\boldsymbol{\lambda}^{\star} = \{\lambda_j^{\star}\}_{j=1}^{L}$ to denote the optimal solution to Problem $(\mathcal{M}')$.

The optimal solution to Problem $(\mathcal{M}')$ is feasible in Problem (B.2)**.** By Lemma 18, if the boundary conditions are binding (i.e., $\sum_{j=1}^{L} \tau \lambda_j^{\star} = \sum_{j=1}^{L} \tau_j \lambda_j^{\dagger}$ or $\sum_{j=1}^{L} \tau \lambda_j^{\star} = \sum_{j=1}^{L} \tau_j \lambda_j^{\ddagger}$), the optimal solution to Problem $(\mathcal{M}')$ is given by a full prioritization policy. Thus, the optimal boundary solutions must be feasible in Problem (B.2).

Second, we want to show that if the boundary conditions are not binding, the optimal solution to Problem $(\mathcal{M}')$ is feasible in Problem (B.2). That is, we want to show that if $\sum_{j=1}^{L} \tau \lambda_j^{\star} > \sum_{j=1}^{L} \tau_j \lambda_j^{\dagger}$ and $\sum_{j=1}^{L} \tau \lambda_j^{\star} < \sum_{j=1}^{L} \tau_j \lambda_j^{\ddagger}$, then $\boldsymbol{\lambda}^{\star}$ is feasible in Problem (B.2).

Without loss of generality, for any $j \in \{1, \ldots, L\}$, we assume that $\mu_j \tau_j \geqslant \mu_{j+1}\tau_{j+1}$. By setting the gradient to be zero and taking the domain $[0, \mu_j)$ into account, the optimal interior solution to Problem $(\mathcal{M}')$ is:

$$\lambda_j^{int} = [\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\sqrt{\frac{\hat{r}}{p-\hat{r}}}]\mathbf{1}_{\mu_j\tau_j \geqslant \hat{r}/(p-\hat{r})} \quad \forall j \in \{1, \ldots, L\} \tag{B.9}$$

Suppose $\boldsymbol{\lambda}^\star = \boldsymbol{\lambda}^{int}$. Note that in this case, p must be larger than $\hat{r}$ and $\mu_1\tau_1 > \hat{r}/(p-\hat{r})$, otherwise the constraint $\sum_{j=1}^L \tau\lambda_j^\star \geqslant \sum_{j=1}^L \tau_j\lambda_j^\dagger$ must bind.

Notice the similarity between Equation (B.9) and Equation (B.12). In the proof of Lemma 15, we know that $\frac{-d_k(N)+\sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}$ is continuous in N. Also, by the definition in Equation (B.12), $\frac{-d_k(N)+\sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}$ converges to 0 as $N \to \infty$; and because of Equation (B.20), $\frac{-d_k(0)+\sqrt{\Delta_k(0)}}{2(\sum_{j=1}^{k(0)}\sqrt{\mu_j\tau_j})} = \sqrt{\mu_1\tau_1}$. Thus, by the intermediate value theorem, there exists $\overline{N}$ such that

$$\frac{-d_k(\overline{N}) + \sqrt{\Delta_k(\overline{N})}}{2(\sum_{j=1}^{k(\overline{N})}\sqrt{\mu_j\tau_j})} = \sqrt{\frac{\hat{r}}{p-\hat{r}}}$$

And we have

$$\hat{r}\sum_{j=1}^{L}\tau_j\lambda_j^\star = h(\overline{N}),\ \lambda_j^\star = \lambda_j^\star(\overline{N})$$

where $\lambda_j^\star(\cdot)$ and $h(\cdot)$ are given by Equation (B.12) and Equation (B.13), respectively.

Since we assume the boundary constraints are not binding, $\sum_{j=1}^L \tau_j\lambda_j^\dagger = h(N_A + N_H^\dagger) < h(\overline{N}) < h(N_A + N_H^\ddagger) = \sum_{j=1}^L \tau_j\lambda_j^\ddagger$ by Lemma 17. It means there exists $N_H^\star = \overline{N} - N_A > 0$ such that $h(\overline{N}) = h(N_A + N_H^\star) = \sum_{j=1}^L \tau_j\lambda_j^\star$, where $\forall j,\ \lambda_j^\star = \lambda_j^\star(N_A + N_H^\star)$, and $N_H^\dagger < N_H^\star < N_H^\ddagger$.

The remaining step is to show that there exists an implementable policy such that $\boldsymbol{\lambda}^\star$ and $N_H^\star$ can be achieved at equilibrium and feasible in Problem (B.2). Given $\boldsymbol{\lambda}^\star$, by Little's law, we need the average number of vehicles at each location j to be $N_j^\star = (\tau_j + \frac{1}{\mu_j - \lambda_j^\star})\lambda_j^\star$. This implies that we can achieve λ_j as long as the policy is non-idling and maintain $N_j^\star$ vehicles at location j in equilibrium. Notice that $\sum_{j=1}^L N_j^\star = N_H^\star + N_A$ because $\lambda_j^\star = \lambda_j^\star(N_A + N_H^\star)$.

Now we need to find a non-idling policy that maintains $N_j^\star$ vehicles at location j and satisfies the wage equilibrium. Here, we propose a randomization approach with two fully-prioritizing policies.

1. π_1: We first fully prioritize AVs by solving Problem (B.4) (i.e. $\lambda_{A,j}^{\pi_1} = \lambda_{A,j}^\dagger$ and $N_{A,j}^{\pi_1} = (\tau_j + \frac{1}{\mu_j - \lambda_{A,j}^\dagger})\lambda_{A,j}^\dagger$). Then, we use the remaining capacity $N_j^\star - N_{A,j}^{\pi_1}$ to allocate HVs. That is, $N_{H,j}^{\pi_1} = N_j^\star - N_{A,j}^{\pi_1}$ and $\lambda_{H,j}^{\pi_1} = \lambda_j^\star - \lambda_{A,j}^{\pi_1}$.

2. π_2: Given $N_H^\star$, we first fully prioritize HVs with the max-arrival-rate allocation. (i.e. $\lambda_{H,j}^{\pi_2} = \lambda_j^\star(N_H^\star)$ and $\sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_2} = h(N_H^\star)$, where $\lambda_j^\star(\cdot)$ and $h(\cdot)$ are given by Equation (B.12) and Equation (B.13), respectively). Because $N_H^\star < N_A + N_H^\star$, we have $\lambda_{H,j}^{\pi_2} < \lambda_j^\star$ and $N_{H,j}^{\pi_2} = (\tau_j + \frac{1}{\mu_j - \lambda_{H,j}^{\pi_2}})\lambda_{H,j}^{\pi_2} < N_j^\star$. Then, we use the remaining capacity $N_j^\star - N_{H,j}^{\pi_2}$ to allocate AVs. That is, $N_{A,j}^{\pi_2} = N_j^\star - N_{H,j}^{\pi_2}$ and $\lambda_{A,j}^{\pi_2} = \lambda_j^\star - \lambda_{H,j}^{\pi_2}$.

The above policies and their random combinations defined by Definition 5 are implementable and must satisfy the constraints of Problem (B.2) except the wage equilibrium Equation (B.2c). Therefore, the last step is to show that $N_H^\star$ is in equilibrium by adopting $\bar{\pi}(\theta)$ for some θ, which is a suitable randomization of π_1 and π_2 defined by Definition 5. Because of Lemma 19, it is sufficient to show that there exists $\omega \in [0, 1]$ such that $rN_H^\star = \gamma P_H \sum_{j=1}^L \tau_j(\omega\lambda_{H,j}^{\pi_1} + (1-\omega)\lambda_{H,j}^{\pi_2})$. To see this, we only need to show $\gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_1} \leqslant rN_H^\star \leqslant \gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_2}$.

1. $\gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_1} \leqslant rN_H^\star$:

   By definition of π_1, $\sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_1} = \sum_{j=1}^L \tau_j(\lambda_j^\star - \lambda_{A,j}^\dagger) = (h(N_A + N_H^\star) - h(N_A))$. And by Problem (B.5) and Lemma 17, $\gamma P_H(h(N_A + N_H^\dagger) - h(N_A)) = rN_H^\dagger$.

   Now there are two cases: $N_H^\dagger = 0$ or $N_H^\dagger > 0$. If $N_H^\dagger = 0$, by Lemma 20 and Lemma 15, $\gamma P_H(h(N_A + N_H) - h(N_A)) \leqslant rN_H$ for all $N_H > 0$. If $N_H^\dagger > 0$ because $N_H^\dagger < N_H^\star$, let

$\alpha = N_H^\dagger / N_H^\star \in (0,1)$. Since $h(N)$ is strictly concave by Lemma 15, we must have:

$$\begin{aligned}
&\alpha h(N_H^\star + N_A) + (1-\alpha)h(N_A) < h(\alpha N_H^\star + N_A) = h(N_H^\dagger + N_A) \\
\implies &\alpha h(N_H^\star + N_A) - \alpha h(N_A) < h(N_H^\dagger + N_A) - h(N_A) \\
\implies &\frac{\alpha h(N_H^\star + N_A) - \alpha h(N_A)}{\alpha N_H^\star} < \frac{h(N_H^\dagger + N_A) - h(N_A)}{\alpha N_H^\star} \\
\implies &\frac{h(N_H^\star + N_A) - h(N_A)}{N_H^\star} < \frac{h(N_H^\dagger + N_A) - h(N_A)}{N_H^\dagger} = r
\end{aligned}$$

Thus, $\gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_1} = \gamma P_H (h(N_A + N_H^\star) - h(N_A)) \leqslant r N_H^\star$.

2. $\gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_2} \geqslant r N_H^\star$:

   Similarly, by definition of π_2, $\sum_{j=1}^L \lambda_{H,j}^{\pi_2} = h(N_H^\star)$. And by Problem (B.6), $\gamma P_H h(N_H^\ddagger) = r N_H^\ddagger$. Because $N_H^\star < N_H^\ddagger$, let $\alpha = N_H^\star / N_H^\ddagger \in (0,1)$. Since $h(N)$ is strictly concave by Lemma 15, we must have:

$$\begin{aligned}
&\alpha h(N_H^\ddagger) + (1-\alpha)h(0) < h(\alpha N_H^\ddagger) = h(N_H^\star) \\
\implies &\alpha h(N_H^\ddagger) < h(N_H^\star) \quad \text{Since } h(0) = 0 \\
\implies &\frac{N_H^\star}{N_H^\ddagger} h(N_H^\ddagger) < h(N_H^\star) \\
\implies &h(N_H^\star) > \frac{r N_H^\star}{\gamma P_H} \quad \text{Since } \gamma P_H h(N_H^\ddagger) = r N_H^\ddagger
\end{aligned}$$

   Thus, $\gamma P_H \sum_{j=1}^L \lambda_{H,j}^{\pi_2} = \gamma P_H h(N_H^\star) \geqslant r N_H^\star$.

Therefore, we are able to find $\omega \in (0,1)$ such that

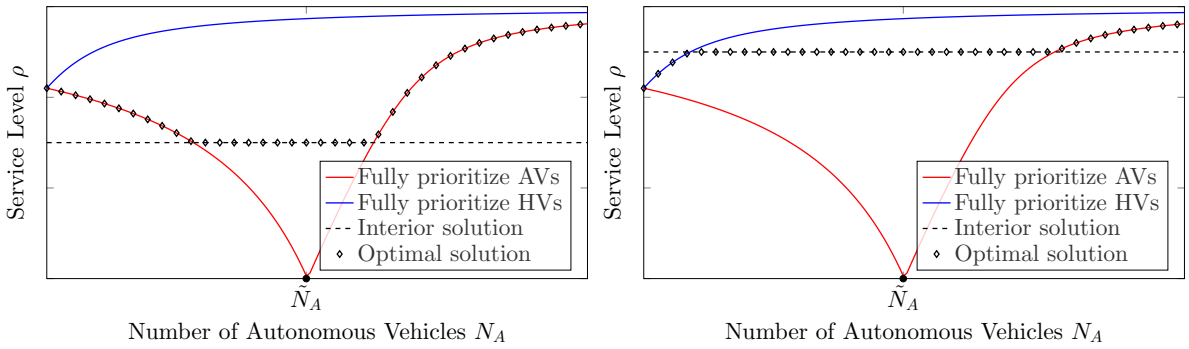
$$r N_H^\star = \omega \gamma p \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_1} + (1-\omega)\gamma p \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_2}$$

And we can get $\omega = \frac{\gamma p \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_2} - r N_H^\star}{\gamma p (\sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_2} - \sum_{j=1}^L \tau_j \lambda_{H,j}^{\pi_1})} \in (0,1)$. Also, according to Lemma 19, we know that there exists a $\theta \in [0,1]$ such that by implementing $\bar{\pi}(\theta)$, we are able to achieve the arrival rates $\omega \lambda_{A,j}^{\pi_1} + (1-\omega)\lambda_{A,j}^{\pi_2}$ and $\omega \lambda_{H,j}^{\pi_1} + (1-\omega)\lambda_{H,j}^{\pi_2}$ at each location j. In other words, we can achieve $\boldsymbol{\lambda}^\star$ and $N_H^\star$ in equilibrium by implementing $\bar{\pi}(\theta)$ for some $\theta \in [0,1]$.

Hence, the optimal solution to Problem ($\mathcal{M}'$) is feasible in Problem (B.2). We conclude that Problem ($\mathcal{M}$) can be reformulated as Problem ($\mathcal{M}'$). □

B.1.3 Proof of the Main Results in Section 3.4 and Section 3.5.

In this section, we prove the main results in Section 3.4 and Section 3.5.



(a) (b)

Figure B.1: The derivation of service levels in Problem ($\mathcal{M}'$). The left plot is an example in which γ is large enough, and we fully prioritize AVs at the beginning; the right plot is an example in which γ is small enough, and we fully prioritize HVs at the beginning.

To help readers understand the reasoning of Theorem 4 and Proposition 8, in Figure B.1, we visualize the service levels with respect to N_A under the different solutions. When Constraint (B.3c) or Constraint (B.3d) of Problem ($\mathcal{M}'$) is binding, the service levels changes as we fully prioritize either AVs or HVs; and when the constraints are not binding, the service level is constant with respect to N_A. Theorem 4 and Proposition 8 show the pattern of service levels under the conditions where prioritizing AVs or prioritizing HVs is optimal at the beginning of introducing AVs.

The following steps require some auxiliary lemmas to complete. Please refer to Appendix B.2.1, Appendix B.2.2 and Appendix B.2.3 for the details.

Proof. Proof of Theorem 4.

By Lemma 25, Constraint (B.3c) is binding if N_A is sufficiently large or if γ is sufficiently large and N_A is sufficiently small. And by Lemma 18, when Constraint (B.3c) is binding, for all $j \in \{1, \ldots, L\}$, $\lambda_j^\star$ (the optimal arrival rate in Problem $(\mathcal{M}')$) is equal to $\lambda_j^\dagger$ (the optimal arrival rate given by Problem (B.5) where we fully prioritize AVs). By Lemma 15, Lemma 17 and Lemma 20, when N_A is sufficiently small, $\lambda_j^\dagger$ is strictly decreasing in N_A, whereas if N_A is sufficiently large, $N_H^\dagger = 0$ and $\lambda_j^\dagger$ is strictly increasing in N_A. In addition, by Equation (B.26) and Equation (B.9), if $\gamma \in (0, 1)$ and N_A in a neighborhood of $\widetilde{N}_A$ where $\widetilde{N}_A$ is defined in Lemma 21, then $\sum_{j=1}^L \tau_j \lambda_j^{int} \geqslant \sum_{j=1}^L \tau_j \lambda_j^\dagger$, so that Constraint (B.3c) is not binding and $\lambda_j^\star = \lambda_j^{int}$ which is constant with respect to N_A as shown in Equation (B.9).

□

Proof. Proof of Proposition 8. By Lemma 25, Constraint (B.3d) is binding when γ and N_A are sufficiently small. And by Lemma 18, when Constraint (B.3d) is binding, for all $j \in \{1, \ldots, L\}$, $\lambda_j^\star$ (the optimal arrival rate of Problem $(\mathcal{M}')$) is equal to $\lambda_j^\ddagger$ (the optimal arrival rate given by Problem (B.7) where we fully prioritize HVs). Thus, when γ and N_A are sufficiently small, it is optimal to prioritize HVs. In addition, when we fully prioritize HVs, $N_A + N_H^\ddagger$ (the total number of vehicles) is strictly increasing in N_A, where $N_H^\ddagger$ is the optimal solution to Problem (B.6). Thus, by Lemma 15 and Lemma 17, $\lambda_j^\ddagger$ is strictly increasing in N_A.

□

Proof. Proof of Proposition 9. The optimization problem in this case is:

$$\max_{N_A \geqslant 0} \left(\begin{array}{ll} \max\limits_{\{\lambda_j\}_{j=1}^L} & \sum\limits_{j=1}^L (p - \hat{r})\tau_j \lambda_j + \hat{r} N_A - \hat{r} \sum\limits_{j=1}^L \dfrac{\lambda_j}{\mu_j - \lambda_j} \\ \text{s.t.} & \{\lambda_j\}_{j=1}^L \in \mathcal{A}, \\ & \lambda_j <\in [0, \mu_j), j \in \{1, \ldots, L\}. \end{array} \right) - C_A N_A \qquad (\mathcal{M}'_{N_A})$$

where $\mathcal{A}$ is the achievable region defined in Section 3.3 and depends on N_A. The problem inside the parenthesis is exactly Problem $(\mathcal{M}')$.

If $C_A \leqslant \hat{r}$, let $\check{\lambda}_j > 0$ and $\check{N}_A \geqslant 0$ denote an optimal solution to Problem $(\mathcal{M}'_{N_A})$. By Little's law, the average number of vehicles is $\sum_{j=1}^{L}(\tau_j + 1/(\mu_j - \check{\lambda}_j))\check{\lambda}_j$, so the average number of HVs is $\check{N}_H = \sum_{j=1}^{L}(\tau_j + 1/(\mu_j - \check{\lambda}_j))\check{\lambda}_j - \check{N}_A$. For the sake of contradiction, suppose $\check{N}_H > 0$. Then, we can find another feasible solution with $\widetilde{N}_A = \check{N}_H + \check{N}_A$ and $\tilde{\lambda}_j = \check{\lambda}_j \ \forall j$ such that the objective value of Problem $(\mathcal{M}'_{N_A})$ is increased by $(\hat{r} - C_A)\check{N}_H$. Thus, it is optimal to only operate AVs if $C_A \leqslant r$.

Now suppose $C_A > \hat{r}$ and let $\hat{N}_A$ denote the threshold of N_A such that all the HVs leave the market. When only AVs are operated, the profit is $ph(N_A) - C_A N_A$ and we know $h(\cdot)$ is strictly concave by Lemma 15. Thus, we only need to show $ph'(\hat{N}_A) \leqslant \hat{r}$ so that the profit will be higher if the number of AVs is lower than $\hat{N}_A$.

For the sake of contradiction, suppose $ph'(\hat{N}_A) > \hat{r}$. By Lemma 20,

$$\begin{aligned} & ph'(\hat{N}_A) = p \cdot \frac{r}{\gamma P_H} > \hat{r} \\ \iff & p \cdot \frac{r}{\gamma P_H} > \frac{r(p - (1-\gamma)P_H)}{\gamma P_H} \\ \iff & p > p - (1-\gamma)P_H \end{aligned}$$

However, this is impossible since $\gamma \in (0, 1)$ and $P_H > 0$. Therefore, it is never optimal to only operate AVs if $C_A > r$.

Lastly, to show a case in which the platform operates both AVs and HVs when $C_A > \hat{r}$, we only need to show a scenario where $ph'(N_H^{\ddagger}) > C_A$ when $N_A = 0$. by Lemma 24, we can see that when $N_A = 0$, the equilibrium number of HVs, $N_H^{\ddagger}$, is constant in p, but $ph'(N_H^{\ddagger})$ is increasing and unbounded with respect to p. Therefore, when $C_A > \hat{r}$, if p is sufficiently large, it is optimal to operate both AVs and HVs.

□

Proof. Proof of Theorem 5. When the boundary constraint of Problem $(\mathcal{M}')$ is binding, by

Lemma 18, we fully prioritize AVs or fully prioritize HVs. And in this case, by Lemma 17 and Lemma 22, we must have $\forall i, j \in \{1, \ldots, L\}$, if $i < j$, $\rho_i^\star \geqslant \rho_j^\star$. Additionally, when the boundary constraints are not binding, by Equation (B.9), the corresponding service level is:

$$\rho_j^\star = [1 - \sqrt{\frac{\hat{r}}{\mu_j \tau_j (p - \hat{r})}}]\mathbf{1}_{\mu_j \tau_j \geqslant \hat{r}/(p-\hat{r})} \quad \forall j \in \{1, \ldots, L\} \tag{B.10}$$

It is easy to see that if $i < j$, $\rho_i^\star \geqslant \rho_j^\star$because $\mu_i \tau_i \geqslant \mu_j \tau_j$. Hence, the service level in high-demand areas is always higher than the service level in low-demand areas.

Second, when the boundary constraints are not binding, by Equation (B.10), $\forall i, j \in \{1, \ldots, L\}$, $\frac{\partial \rho_i^\star}{\partial N_A} = \frac{\partial \rho_j^\star}{\partial N_A} = 0$. Additionally, if the boundary constraint of Problem $(\mathcal{M}')$ is binding, by Lemma 18, we fully prioritize AVs or fully prioritize HVs. When we fully prioritize HVs, since $N_H^\ddagger$ is irrelevant with N_A, $\frac{\partial N_A + N_H^\ddagger}{\partial N_A} = \frac{\partial N_A}{\partial N_A} = 1 > 0$, where $N_H^\ddagger$ is the optimal solution to Problem (B.6). Thus, by Lemma 17 and Lemma 22, $\forall i < j$, $\frac{\partial \rho_i^\star}{\partial N_A} \leqslant \frac{\partial \rho_j^\star}{\partial N_A}$. When we fully prioritize AVs, by Lemma 20, $\frac{\partial N_A + N_H^\dagger}{\partial N_A}$ is either negative or positive, where $N_H^\dagger$ is the optimal solution to Problem (B.5). Thus, by Lemma 17 and Lemma 22, we have $\forall i < j, |\frac{\partial \rho_i^\star}{\partial N_A}| \leqslant |\frac{\partial \rho_j^\star}{\partial N_A}|$. Hence, $|\frac{\partial \rho_j^\star}{\partial N_A}| \leqslant |\frac{\partial \rho_{j+1}^\star}{\partial N_A}|$.

□

Proof. Proof of Proposition 10.

In Theorem 4, we showed that it is optimal to fully prioritize AVs, when γ is high enough and N_A is low enough. In the following, let us suppose the platform is fully prioritizing AVs.

At location j, let $N_{A,j}(N_A)$ denote the average number of AVs, $N_{H,j}(N_A + N_H^\dagger)$ denote the average number of HVs and $N_j(N_A + N_H^\dagger) = N_{A,j}(N_A) + N_{H,j}(N_A + N_H^\dagger)$ denote the average number of vehicles, where $N_H^\dagger$ is the optimal solution to Problem (B.5).

For the concentration of AVs, by Lemma 17 and Lemma 23,

$$\frac{\partial N_{A,j-1}(N_A)}{\partial N_A} \geqslant \frac{\partial N_{A,j}(N_A)}{\partial N_A} \geqslant 0$$

Therefore, more AVs will concentrate in high-demand areas.

Second, for the concentration of HVs, by Lemma 20, when we fully prioritize AVs and there are HVs in equilibrium, $\frac{\partial(N_A+N_H^\dagger)}{\partial N_A} < 0$. Thus, by Lemma 17 and Lemma 23, we have:

$$\frac{\partial N_{j-1}(N_A+N_H^\dagger)}{\partial N_A} \leqslant \frac{\partial N_j(N_A+N_H^\dagger)}{\partial N_A} \leqslant 0$$

Because $N_j(N_A+N_H^\dagger) = N_{H,j}(N_A+N_H^\dagger) + N_{A,j}(N_A)$,

$$\begin{aligned}
&\frac{\partial N_{j-1}(N_A+N_H^\dagger)}{\partial N_A} \leqslant \frac{\partial N_j(N_A+N_H^\dagger)}{\partial N_A} \leqslant 0 \\
\implies &\frac{\partial N_{H,j-1}(N_A+N_H^\dagger)}{\partial N_A} + \frac{\partial N_{A,j-1}(N_A)}{\partial N_A} \leqslant \frac{\partial N_{H,j}(N_A+N_H^\dagger)}{\partial N_A} + \frac{\partial N_{A,j}(N_A)}{\partial N_A} \leqslant 0 \\
\implies &\frac{\partial N_{H,j-1}(N_A+N_H^\dagger)}{\partial N_A} \leqslant \frac{\partial N_{H,j}(N_A+N_H^\dagger)}{\partial N_A} + \frac{\partial N_{A,j}(N_A)}{\partial N_A} - \frac{\partial N_{A,j-1}(N_A)}{\partial N_A} \leqslant 0 \\
\implies &\frac{\partial N_{H,j-1}(N_A+N_H^\dagger)}{\partial N_A} \leqslant \frac{\partial N_{H,j}(N_A+N_H^\dagger)}{\partial N_A} \leqslant 0
\end{aligned}$$

Therefore, more HVs will leave high-demand areas.

□

Proof. Proof of Proposition 11. By Lemma 24, when $N_A = 0$, the service level at j can be expressed as

$$\rho_j^\ddagger = \left[1 - \sqrt{\frac{1}{\mu_j\tau_j}} \cdot \frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1}^J \sqrt{\mu_j\tau_j})}\right] \mathbf{1}_{j\leqslant J}$$

where $d = \sum_{j=1}^J (1-(1-\gamma P_H/r)\mu_j\tau_j)$, $\Delta = d^2 + 4(1-\gamma P_H/r)(\sum_{j=1}^J \sqrt{\mu_j\tau_j})^2$, and J is also defined in Lemma 24. By Theorem 4, the minimum service level can be derived by the interior solution in Equation (B.9):

$$\rho_j^{int} = \left[1 - \sqrt{\frac{1}{\mu_j\tau_j}}\sqrt{\frac{\hat{r}}{p-\hat{r}}}\right] \mathbf{1}_{\mu_j\tau_j \geqslant r/(p-\hat{r})} \quad \forall j \in \{1,\ldots,L\}$$

It implies that the maximum loss of service level is

$$\Delta\rho_j = \rho_j^\ddagger - \rho_j^{int} = \sqrt{\frac{1}{\mu_j\tau_j}}\left[\sqrt{\frac{\hat{r}}{p-\hat{r}}} - \frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1}^J \sqrt{\mu_j\tau_j})}\right]$$

Thus, $\Delta\rho_j > \Delta\rho_i$ since $\mu_i\tau_i > \mu_j\tau_j$

In particular,

$$\begin{aligned}\frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})} &= \frac{1}{2(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})}\cdot\frac{\Delta-d^2}{\sqrt{\Delta}+d}\\ &= \frac{1}{2(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})}\cdot\frac{4(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})^2}{\sqrt{\Delta}+d}\\ &= \frac{2(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})}{\sqrt{\Delta}+d}\\ &\leqslant \frac{2(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})}{d}\end{aligned}$$

By L'Hôpital's rule,

$$\lim_{\mu_i\tau_i\to\infty}\frac{2(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})}{d} = \lim_{\mu_i\tau_i\to\infty}\frac{1}{(\gamma P_H/r-1)\sqrt{\mu_i\tau_i}} = 0$$

Because of the squeeze theorem,

$$\lim_{\mu_i\tau_i\to\infty}\frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})} = 0$$

Therefore,

$$\lim_{\mu_i\tau_i\to\infty}\Delta\rho_j = \sqrt{\frac{1}{\mu_j\tau_j}}\left[\sqrt{\frac{r}{p-\hat{r}}}-0\right] = \sqrt{\frac{\hat{r}}{\mu_j\tau_j(p-\hat{r})}}$$

$$\lim_{\mu_i\tau_i\to\infty}\Delta\rho_i = \lim_{\mu_i\tau_i\to\infty}\sqrt{\frac{1}{\mu_i\tau_i}}\left[\sqrt{\frac{\hat{r}}{p-\hat{r}}}-0\right] = 0$$

□

B.2 Supporting Material for Proofs

B.2.1 Supporting Material: the maximum arrival rate function in a single-type system.

To facilitate the analysis for Problem ($\mathcal{M}$), we define the single-type maximum arrival rate function $h(N) : \mathbb{R}_+ \to \mathbb{R}_+$ as the optimal value of

$$\begin{aligned} \max_{\lambda_j, j\in\{1,\ldots,L\}} \quad & \sum_{j=1}^{L} \cdot \tau_j \cdot \lambda_j \\ \text{s.t.} \quad & \sum_{j=1}^{L} \left(\tau_j + \frac{1}{\mu_j - \lambda_j}\right) \cdot \lambda_j = N, \\ & \lambda_j \in [0, \mu_j), \qquad\qquad j \in \{1, \ldots, L\}, \end{aligned} \tag{B.11}$$

where λ_j is the arrival rate of the single vehicle type at location j.

In the queueing model, the optimal solution, $\lambda_j^\star(N)$, and the optimal value, $h(N)$, of Problem (B.11) are important for helping us complete the analysis. In this subsection, we solve Problem (B.11) and analyze the properties of $h(N)$ as a preparation for the other results. Without loss of generality, we assume for any $j \in \{1, \ldots, L\}$, $\mu_j\tau_j \geqslant \mu_{j+1}\tau_{j+1}$.

The following Lemma 14 solves Problem (B.11) and illustrates that when we increase the fleet size, locations with higher demand are served first. That is, for each location j, there exists a threshold $\overline{N}_j \geqslant 0$ such that the average number of vehicles at location j is positive if and only if $N > \overline{N}_j$. Also, $\overline{N}_j \leqslant \overline{N}_{j+1}$ because $\mu_j\tau_j \geqslant \mu_{j+1}\tau_{j+1}$.

Lemma 14 (Solution to Problem (B.11))**.** For any $N > 0$, let $k(N) = \max_{k\in\{1,\ldots,L\}}\{k|N > \overline{N}_k\}$, where $\{\overline{N}_j\}_{j=1}^L$ is a non-decreasing sequence of constants that are irrelevant to N and derived by the other parameters in Problem (B.11). Then we can express the optimal solution as:

$$\lambda_j^\star(N) = [\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\frac{-d_k(N) + \sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}]\mathbf{1}_{j\leqslant k(N)} \tag{B.12}$$

where $d_k(N) = N + k(N) - \sum_{j=1}^{k(N)} \mu_j \tau_j$, $\Delta_k(N) = d_k(N)^2 + 4(\sum_{j=1}^{k(N)} \sqrt{\mu_j \tau_j})^2$, and $\mathbf{1}_{j \leqslant k(N)}$ is a binary indicator function.

And the optimal objective value $h(N)$ can be expressed as:

$$h(N) = \sum_{j=1}^{L} \tau_j \lambda_j^\star(N) = \frac{1}{2}[N + \sum_{j=1}^{k(N)} \tau_j \mu_j + k(N) - \sqrt{\Delta_k(N)}] \tag{B.13}$$

Proof. Proof of Lemma 14. We can use the method of Lagrange multipliers to solve Problem (B.11) and construct the Lagrangian function:

$$\mathcal{L}(\{\lambda_j\}_{j=1}^L, \theta, \{\phi_j\}_{j=1}^L) = -\sum_{j=1}^{L} \tau_j \lambda_j + \theta(\sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j})\lambda_j - N) - \sum_{j=1}^{L} \phi_j \lambda_j$$

where θ and ϕ_j are Lagrange multipliers. The Kuhn-Tucker conditions are:

- Stationarity:

$$-\tau_j + \theta(\tau_j + \frac{\mu_j}{(\mu_j - \lambda_j^\star)^2}) - \phi_j = 0$$

- Primal feasibility:

$$\sum_{j=1}^{L} \left(\tau_j + \frac{1}{\mu_j - \lambda_j}\right) \cdot \lambda_j = N$$

$$\lambda_j \geqslant 0, j \in \{1, \ldots, L\}$$

- Dual feasibility:

$$\phi_j \geqslant 0, j \in \{1, \ldots, L\}$$

- Complementary slackness:

$$\phi_j \lambda_j = 0, j \in \{1, \ldots, L\}$$

In the following, let us use a superscript $\star$ to denote an optimal solution. Because of the complementary slackness, we consider two cases: either $\lambda_j^\star > 0$ or $\lambda_j^\star = 0$. First, if $\lambda_j^\star$ is positive at some location j, then $\phi_j^\star = 0$, and the optimal $\lambda_j^\star$ and $\theta^\star$ must satisfy:

$$\tau_j - \theta^\star(\tau_j + \frac{\mu_j}{(\mu_j - \lambda_j^\star)^2}) = 0$$

And we have:

$$(\mu_j - \lambda_j^\star)^2 = \frac{\theta^\star \mu_j}{(1-\theta^\star)\tau_j}$$

$$\lambda_j^\star = \mu_j - \sqrt{\frac{\theta^\star \mu_j}{(1-\theta^\star)\tau_j}} \quad \text{because } \lambda_j^\star < \mu_j$$

Notice that this implies that $\theta^\star$ and $1-\theta^\star$ must be positive (i.e. $\theta^\star \in (0,1)$).

Now by considering the equality constraint:

$$\sum_{j=1,\lambda_j^\star>0}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j^\star})\lambda_j^\star = N$$

$$\implies \sqrt{\frac{\theta^\star}{1-\theta^\star}} = \frac{-d+\sqrt{\Delta}}{2(\sum_{j=1,\lambda_j^\star>0}^{L}\sqrt{\mu_j\tau_j})} \tag{B.14}$$

where $d = N + \sum_{j=1,\lambda_j^\star>0}^{L}(1-\mu_j\tau_j)$ and $\Delta = d^2 + 4(\sum_{j=1,\lambda_j^\star>0}^{L}\sqrt{\mu_j\tau_j})^2$. Notice that $\sqrt{\frac{\theta^\star}{1-\theta^\star}} \neq \frac{-d-\sqrt{\Delta}}{2(\sum_{j=1,\lambda_j^\star>0}^{L}\sqrt{\mu_j\tau_j})}$, because $d < \sqrt{\Delta}$ and $\sqrt{\frac{\theta^\star}{1-\theta^\star}} > 0$. Therefore, when $\lambda_j^\star > 0$, we must have:

$$\lambda_j^\star = \mu_j - \sqrt{\frac{\mu_j}{\tau_j}} \cdot \frac{-d+\sqrt{\Delta}}{2(\sum_{j=1,\lambda_j^\star>0}^{L}\sqrt{\mu_j\tau_j})}$$

Second, if $\lambda_j^\star$ is zero at some location j, by the Kuhn-Tucker conditions, we have $\phi^\star \geqslant 0$ and:

$$\tau_j - \theta^\star(\tau_j + \frac{1}{\mu_j}) \leqslant 0$$

$$\iff \mu_j\tau_j \leqslant \frac{\theta^\star}{1-\theta^\star} \tag{B.15}$$

This implies that for any i and j, if $\mu_j\tau_j \geqslant \mu_i\tau_i$ and $\lambda_j^\star = 0$, then $\lambda_i^\star$ must be also zero. Additionally, because $d < \sqrt{\Delta}$:

$$\frac{\partial(-d+\sqrt{\Delta})}{\partial N} = \frac{d}{\sqrt{\Delta}} - 1 < 0 \tag{B.16}$$

And we have:

$$\lim_{N\to\infty}(-d+\sqrt{\Delta}) = \lim_{N\to\infty}\frac{\Delta - d^2}{d+\sqrt{\Delta}} \to 0 \tag{B.17}$$

Thus, $\sqrt{\frac{\theta^\star}{1-\theta^\star}}$ decreases in N and converges to 0. This means that as we increase the value of N, $\lambda_j^\star$ must change from zero to a positive value.

Recall that we assumed for any $j \in \{1, \ldots, L-1\}$, $\mu_j\tau_j \geqslant \mu_{j+1}\tau_{j+1}$. Then, the above analysis implies that there exists a non-decreasing sequence $\{\overline{N}_j\}_{j=1}^L$, $\overline{N}_j \leqslant \overline{N}_{j+1}$ such that $\forall N \leqslant \overline{N}_j,\ \lambda_j^\star = 0$, and $\forall N > \overline{N}_j,\ \lambda_j^\star > 0$. Specifically, by Equation (B.14) and Inequality (B.15), $\forall k \geqslant 1$, $\overline{N}_k$ must be the solution to

$$\frac{-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)}}{2(\sum_{j=1}^k \sqrt{\mu_j\tau_j})} = \sqrt{\mu_k\tau_k} \tag{B.18}$$

where $d_k(\overline{N}_k) = \overline{N}_k + k - \sum_{j=1}^k \mu_j\tau_j$ and $\Delta_k(\overline{N}_k) = d_k(\overline{N}_k)^2 + 4(\sum_{j=1}^k \sqrt{\mu_j\tau_j})^2$. In addition, we can see that

$$\begin{aligned}
\frac{-d_1(0) + \sqrt{\Delta_1(0)}}{2\sqrt{\mu_1\tau_1}} &= \frac{-(1-\mu_1\tau_1) + \sqrt{(1-\mu_1\tau_1)^2 + 4\mu_1\tau_1}}{2\sqrt{\mu_1\tau_1}} \\
&= \frac{\mu_1\tau_1 - 1 + \sqrt{(1+\mu_1\tau_1)^2}}{2\sqrt{\mu_1\tau_1}} \\
&= \frac{2\mu_1\tau_1}{2\sqrt{\mu_1\tau_1}} = \sqrt{\mu_1\tau_1}
\end{aligned}$$

Thus, $\overline{N}_1 = 0$. This corresponds to the fact that $\lambda_1^\star(N)$ has to be positive for any $N > 0$; otherwise, all of $\lambda_j^\star(N)$ will be zero and Little's law will be violated.

Then, for any N, let $k(N) = \max\{k | N > \overline{N}_k\}$, we can express the optimal solution as:

$$\lambda_j^\star(N) = \left[\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\frac{-d_k(N) + \sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)} \sqrt{\mu_j\tau_j})}\right] \mathbf{1}_{j \leqslant k(N)}$$

where $d_k(N) = N + k(N) - \sum_{j=1}^{k(N)} \mu_j\tau_j$ and $\Delta_k(N) = d_k(N)^2 + 4(\sum_{j=1}^{k(N)} \sqrt{\mu_j\tau_j})^2$.

And the optimal objective value $h(N)$ can be expressed as:

$$h(N) = \sum_{j=1}^L \tau_j\lambda_j^\star(N) = \frac{1}{2}\left[N + \sum_{j=1}^{k(N)} \tau_j\mu_j + k(N) - \sqrt{\Delta_k(N)}\right]$$

□

The following Lemma 15 analyzes the properties of $h(N)$ given by Equation (B.13), which helps us to derive the other results.

Lemma 15 (Properties of $h(N)$)**.** $h(N)$ in Equation (B.13) is continuous, strictly increasing, differentiable, and strictly concave. In addition, $\forall j \in \{1, \ldots, L\}$, $\lambda_j^\star(N)$ in Equation (B.12) is continuous for $N \geqslant 0$ and strictly increasing for $N \geqslant \overline{N}_j$.

Proof. Proof of Lemma 15. Let us prove the properties of $h(N)$ in order.

Continuity of $h(N)$ First, we want to show $h(N)$ is continuous at any $N \geqslant 0$. In fact, we only need to show $\lambda_j^\star(N)$ is continuous at all $N \in \{\overline{N}_j\}_{j=1}^L$, where $\{\overline{N}_j\}_{j=1}^L$ can be derived by Equation (B.18). Let $k \in \{1, \ldots, L\}$, we want to show $\forall j \in \{1, \ldots, L\}$, $\lim_{N \to \overline{N}_k^-} \lambda_j^\star(N) = \lim_{N \to \overline{N}_k^+} \lambda_j^\star(N)$.

If $\mu_j \tau_j < \mu_k \tau_k$, then it is clear that $\lim_{N \to \overline{N}_k^-} \lambda_j^\star(N) = \lim_{N \to \overline{N}_k^+} \lambda_j^\star(N) = 0$ by Equation (B.12).

If $\mu_j \tau_j = \mu_k \tau_k$, by Equation (B.18), we have

$$\lim_{N \to \overline{N}_k^+} \lambda_j^\star(N) = \mu_j - \sqrt{\frac{\mu_j}{\tau_j}} \frac{-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)}}{2(\sum_{j=1}^k \sqrt{\mu_j \tau_j})} = \mu_j - \sqrt{\frac{\mu_j}{\tau_j}} \sqrt{\mu_k \tau_k} = 0$$

And it is clear that $\lim_{N \to \overline{N}_k^-} \lambda_j^\star(N) = 0$ by Equation (B.12). Thus, we also have $\lim_{N \to \overline{N}_k^-} \lambda_j^\star(N) = \lim_{N \to \overline{N}_k^+} \lambda_j^\star(N) = 0$.

If $\mu_j \tau_j > \mu_k \tau_k$, we have

$$\lim_{N \to \overline{N}_k^-} \lambda_j^\star(N) = \mu_j - \sqrt{\frac{\mu_j}{\tau_j}} \frac{-d_{k-1}(\overline{N}_k) + \sqrt{\Delta_{k-1}(\overline{N}_k)}}{2(\sum_{j=1}^{k-1} \sqrt{\mu_j \tau_j})}$$

$$\lim_{N \to \overline{N}_k^+} \lambda_j^\star(N) = \mu_j - \sqrt{\frac{\mu_j}{\tau_j}} \frac{-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)}}{2(\sum_{j=1}^{k} \sqrt{\mu_j \tau_j})}$$

We want to show $\frac{-d_{k-1}(\overline{N}_k) + \sqrt{\Delta_{k-1}(\overline{N}_k)}}{2(\sum_{j=1}^{k-1} \sqrt{\mu_j \tau_j})} = \frac{-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)}}{2(\sum_{j=1}^{k} \sqrt{\mu_j \tau_j})} = \sqrt{\mu_k \tau_k}$:

Start with Equation (B.18):

$$\frac{-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)}}{2(\sum_{j=1}^{k}\sqrt{\mu_j\tau_j})} = \sqrt{\mu_k\tau_k}$$

$$-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)} = 2(\sum_{j=1}^{k}\sqrt{\mu_j\tau_j})\sqrt{\mu_k\tau_k}$$

$$\sqrt{\Delta_k(\overline{N}_k)} = d_k(\overline{N}_k) + 2(\sum_{j=1}^{k}\sqrt{\mu_j\tau_j})\sqrt{\mu_k\tau_k} \tag{B.19a}$$

$$\Delta_k(\overline{N}_k) = d_k(\overline{N}_k)^2 + 4\mu_k\tau_k(\sum_{j=1}^{k}\sqrt{\mu_j\tau_j})^2 + 4\sqrt{\mu_k\tau_k}d_k(\overline{N}_k)(\sum_{j=1}^{k}\sqrt{\mu_j\tau_j})$$

$$\sum_{j=1}^{k}\sqrt{\mu_j\tau_j} = \mu_k\tau_k(\sum_{j=1}^{k}\sqrt{\mu_j\tau_j}) + d_k(\overline{N}_k)\sqrt{\mu_k\tau_k} \tag{B.19b}$$

$$\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j} + \sqrt{\mu_k\tau_k} = \mu_k\tau_k(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j}) + d_{k-1}(\overline{N}_k)\sqrt{\mu_k\tau_k} + \sqrt{\mu_k\tau_k}$$

$$\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j} = \mu_k\tau_k(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j}) + d_{k-1}(\overline{N}_k)\sqrt{\mu_k\tau_k}$$

$$\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j} = \mu_k\tau_k(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j}) + d_{k-1}(\overline{N}_k)\sqrt{\mu_k\tau_k}$$

$$\Delta_{k-1}(\overline{N}_k) = d_{k-1}(\overline{N}_k)^2 + 4\mu_k\tau_k(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j})^2 + 4\sqrt{\mu_k\tau_k}d_{k-1}(\overline{N}_k)(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j})$$

$$-d_{k-1}(\overline{N}_k) + \sqrt{\Delta_{k-1}(\overline{N}_k)} = 2(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j})\sqrt{\mu_k\tau_k}$$

$$\frac{-d_{k-1}(\overline{N}_k) + \sqrt{\Delta_{k-1}(\overline{N}_k)}}{2(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j})} = \sqrt{\mu_k\tau_k}$$

Thus,

$$\frac{-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)}}{2(\sum_{j=1}^{k-1}\sqrt{\mu_j\tau_j})} = \frac{-d_{k-1}(\overline{N}_k) + \sqrt{\Delta_{k-1}(\overline{N}_k)}}{2(\sum_{j=1}^{k}\sqrt{\mu_j\tau_j})} = \sqrt{\mu_k\tau_k} \tag{B.20}$$

Hence, for any $j \in \{1, \ldots, L\}$, $\lambda_j(N)$ is continuous at any $N \geqslant 0$. This implies that $h(N)$ is continuous at any $N \geqslant 0$.

Monotonicity of $h(N)$ Because of Equation (B.12) and Equation (B.16), $\frac{\partial \lambda_j^\star(N)}{\partial N} > 0$ for $N \geqslant \overline{N}_j$. And we know $\overline{N}_1 = 0$, so $\lambda_1^\star(N)$ is strictly increasing for all $N \geqslant 0$. Therefore, $h(N)$ is strictly increasing in $N \geqslant 0$.

Differentiability and Concavity We want to show that $h'(N) \triangleq \frac{\partial h(N)}{\partial N}$ is continuous and strictly decreasing in $N \geqslant 0$. By Equation (B.13), we can get for any $N \neq \overline{N}_k$, $\forall k \in \{1, \ldots, L\}$:

$$h'(N) = \frac{1}{2}(1 - \frac{d_k(N)}{\sqrt{\Delta_k(N)}}) \tag{B.21}$$

And the second derivative $h''(N)$ is

$$h''(N) = \frac{d_k(N)^2 - \Delta_k(N)}{\sqrt{\Delta_k(N)}\Delta_k(N)} < 0$$

where the negativity is because of $d_k(N)^2 < \Delta_k(N)$. Thus, $h'(N)$ is strictly decreasing in $N \in (\overline{N}_k, \overline{N}_{k+1})$ for any $k \in \{1, \cdots L-1\}$, or any $N \in (\overline{N}_L, \infty)$. The remaining step is to show $h'(N)$ is continuous at $\{\overline{N}_j\}_{j=1}^L$. Because $\lim_{N \to \overline{N}_k^-} h'(N) = \frac{1}{2}(1 - \frac{d_{k-1}(\overline{N}_k)}{\sqrt{\Delta_{k-1}(\overline{N}_k)}})$ and $\lim_{N \to \overline{N}_k^+} h'(N) = \frac{1}{2}(1 - \frac{d_k(\overline{N}_k)}{\sqrt{\Delta_k(\overline{N}_k)}})$, we only need to show $\frac{d_{k-1}(\overline{N}_k)}{\sqrt{\Delta_{k-1}(\overline{N}_k)}} = \frac{d_k(\overline{N}_k)}{\sqrt{\Delta_k(\overline{N}_k)}}$. By Equation (B.19b):

$$\frac{-d_k(\overline{N}_k) + \sqrt{\Delta_k(\overline{N}_k)}}{2(\sum_{j=1}^k \sqrt{\mu_j \tau_j})} = \sqrt{\mu_k \tau_k} \implies d_k(\overline{N}_k) = \frac{(1 - \mu_k\tau_k)\sum_{j=1}^k \sqrt{\mu_j\tau_j}}{\sqrt{\mu_k\tau_k}}$$

Substitute the above equation into Equation (B.19a), we get:

$$\begin{aligned}\sqrt{\Delta_k(\overline{N}_k)} &= d_k(\overline{N}_k) + 2(\sum_{j=1}^k \sqrt{\mu_j\tau_j})\sqrt{\mu_k\tau_k} \\ &= \frac{(1 + \mu_k\tau_k)\sum_{j=1}^k \sqrt{\mu_j\tau_j}}{\sqrt{\mu_k\tau_k}}\end{aligned}$$

Thus,

$$\frac{d_k(\overline{N}_k)}{\sqrt{\Delta_k(\overline{N}_k)}} = \frac{1 - \mu_k\tau_k}{1 + \mu_k\tau_k}$$

Similarly, because of Equation (B.20), we can repeat the above steps and obtain:

$$\frac{d_{k-1}(\overline{N}_k)}{\sqrt{\Delta_{k-1}(\overline{N}_k)}} = \frac{1-\mu_k\tau_k}{1+\mu_k\tau_k}$$

Thus,

$$\frac{d_{k-1}(\overline{N}_k)}{\sqrt{\Delta_{k-1}(\overline{N}_k)}} = \frac{d_k(\overline{N}_k)}{\sqrt{\Delta_k(\overline{N}_k)}} = \frac{1-\mu_k\tau_k}{1+\mu_k\tau_k}$$

This implies that $h'(N)$ is continuous and strictly decreasing in N. □

B.2.2 Supporting Material for Proposition 7.

Lemma 16 (Full prioritization maximizes arrival rates)**.** With the constraints of Problem $(\mathcal{M})$, $\sum_{j=1}^{L}\tau_j\lambda_{A,j}$ is maximized only if we fully prioritize AVs, and $\sum_{j=1}^{L}\tau_j\lambda_{H,j}$ is maximized only if we fully prioritize HVs.

Proof. Proof of Lemma 16. The proof is similar to the proof of Lemma 2. In fact, a policy that does not fully prioritize AVs (or HVs) can be seen as an idling policy for AVs (or HVs). For instance, in the view of human drivers, no matter whether a policy rejects a request or matches a request with AVs, such a policy does not always match requests with HVs when HVs are available, so this policy idles HVs sometimes. In this sense, the idea in the proof of Lemma 2 also works in Lemma 16. Nonetheless, we formally present the proof of Lemma 16 as the following.

First, we want to show that given fixed N_A (or N_H) and a policy π which does not always prioritize AVs (or HVs), we can always increase the arrival rate of AVs (or HVs) by using a policy $\pi^\star$ which fully prioritizes AVs (or HVs). Here, we only present the case for AVs, and the proof for HVs is exactly the same. Let $N_A^\pi(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$ be the average number of AVs in a system with two types of vehicles under a policy π, where the arrival rates of vehicles are $\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H$ for AVs and HVs respectively. Suppose π does not fully prioritize AVs, so there exists some location j where AVs are not always prioritized. And let $\pi^\star$ denote a policy that always prioritizes AVs at all the locations.

We argue that $N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) > N_A^{\pi^\star}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$. Indeed, given arrival rates to the system, $\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H$, $\pi^\star$ always matches requests with AVs as long as there exist AVs available in the queue, while π might match request with HVs at some location j even if AVs are available. The matching with HVs in π increases the queue size of AVs in location j. This implies that the average number of AVs in the system under π is larger than that under $\pi^\star$. Moreover, for any policy π and location j, we know that $N_{A,j}^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$ increases with $\lambda_{A,j}$.

Given the constraints $N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) = N_A$, there must exist $\lambda_{A,j}^\star > \lambda_{A,j}$ at location j such that $N_A^{\pi^\star}(\boldsymbol{\lambda}_A^\star, \boldsymbol{\lambda}_H^\star) = N_A$, because $N_A^{\pi^\star}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) < N_A^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H) = N_A$ and $N_{A,j}^{\pi}(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_H)$ increases with $\lambda_{A,j}$. Thus, given a constant N_A, $R_A = (P_A - c_A)\sum_{j=1}^{L} \tau_j \lambda_{A,j}$ is maximized only if AVs are fully prioritized. Similarly, given a constant N_H, $R_H = P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j}$ is maximized only if HVs are fully prioritized.

Second, by the above analysis, we know that there exists a policy $\pi^\star$ which fully prioritizes HVs and can produce a higher arrival rate of HVs than a policy π that does not fully prioritize HVs, but $\pi^\star$ might not satisfy the wage equilibrium. In other words, assume π does not always prioritize HVs and is feasible in Problem $(\mathcal{M})$ so that the wage equilibrium is satisfied (i.e. $\gamma \sum_{j=1}^{L} P_H \tau_j \lambda_{H,j} = r N_H$, where N_H is the average number of HVs under π.), we need to show that there exists an equilibrium number of HVs, $N_H^\star$, under a policy $\pi^\star$ that fully prioritizes HVs, such that $N_H^\star \geqslant N_H$.

By the first part, we have:

$$\gamma P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j}^\star > \gamma P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j} = r N_H \tag{B.22}$$

This means that more HVs will enter the market. And by Little's law:

$$\sum_{j=1}^{L} (\tau_j + W_{H,j}^{\pi^\star}(\lambda_{A,j}^\star, \lambda_{H,j}^\star))\lambda_{H,j}^\star = N_H$$

Because $\pi^\star$ fully prioritizes HVs, the expected waiting function of the HV queueing at each

location j is $W^{\pi^\star}_{H,j}(\lambda_{A,j}, \lambda_{H,j}) = 1/(\mu_j - \lambda_{H,j})$, so the above Little's law becomes:

$$\sum_{j=1}^{L}(\tau_j + 1/(\mu_j - \lambda^\star_{H,j}))\lambda^\star_{H,j} = N_H$$

Substitute N_H in Equation (B.22) with the above equation, we have

$$\sum_{j=1}^{L}(\tau_j + 1/(\mu_j - \lambda^\star_{H,j}))\lambda^\star_{H,j} - \frac{\gamma P_H}{r}\sum_{j=1}^{L}\tau_j\lambda^\star_{H,j} < 0$$

The left-hand side of the above inequality is continuous in $\lambda^\star_{H,j} \in (0, \mu_j)$ and converges to ∞ as $\lambda^\star_{H,j} \to \mu_j$. Thus, by the intermediate value theorem, we must be able to find $\tilde{\lambda}_{H,j} \geqslant \lambda^\star_{H,j}$ and $N^\star_H > N_H$ such that

$$\sum_{j=1}^{L}(\tau_j + 1/(\mu_j - \tilde{\lambda}_{H,j}))\tilde{\lambda}_{H,j} - \frac{\gamma P_H}{r}\sum_{j=1}^{L}\tau_j\tilde{\lambda}_{H,j} = 0, \quad \gamma\sum_{j=1}^{L}P_H\tau_j\tilde{\lambda}_{H,j} = rN^\star_H$$

In this case, both Little's law and the wage equilibrium are satisfied. Thus, with the wage equilibrium, $R_H = P_H\sum_{j=1}^{L}\tau_j\lambda_{H,j}$ is maximized only if HVs are fully prioritized. □

The next auxiliary lemma characterizes the optimal solution and objective value when we fully prioritize AVs or HVs. We want to show that if we fully prioritize AVs or HVs, the allocation of vehicles must maximize the overall arrival rates and be consistent with $\lambda^\star_j(\cdot)$ of Equation (B.12) and $h(\cdot)$ of Equation (B.13). Notice that the following $\boldsymbol{\lambda}^\dagger, \boldsymbol{\lambda}^\dagger_A, N^\dagger_H$ are defined in Problem (B.4) and Problem (B.5), which represent the optimal solution when we fully prioritize AVs; $\boldsymbol{\lambda}^\ddagger, \boldsymbol{\lambda}^\ddagger_H, N^\ddagger_H$ are defined in Problem (B.6) and Problem (B.7), which represent the optimal solution when we fully prioritize HVs.

Lemma 17 (Optimal solution to the full prioritization problems)**.** Let $N^\dagger_H$ (or $N^\ddagger_H$) denote the optimal equilibrium number of HVs when we fully prioritize AVs (or HVs), then the optimal overall arrival rate is $\lambda^\star_j(N_A + N^\dagger_H)$ (or $\lambda^\star_j(N_A + N^\ddagger_H)$) and the optimal total arrival rate is $h(N_A + N^\dagger_H)$ (or $h(N_A + N^\ddagger_H)$), where $\lambda^\star(\cdot)$ and $h(\cdot)$ are defined in Equation (B.12) and Equation (B.13). That is, $h(N_A + N^\dagger_H) = \sum_{j=1}^{L}\tau_j\lambda^\dagger_j$, $h(N_A + N^\ddagger_H) = \sum_{j=1}^{L}\tau_j\lambda^\ddagger_j$ where $\forall j \in \{1, \ldots, L\}$, $\lambda^\dagger_j = \lambda^\star_j(N_A + N^\dagger_H)$, $\lambda^\ddagger_j = \lambda^\star_j(N_A + N^\ddagger_H)$.

Proof. Proof of Lemma 17. We first want to find the expressions of $\boldsymbol{\lambda}^\dagger$ and $\boldsymbol{\lambda}^\ddagger$. Let us start with the case where we fully prioritize AVs.

Clearly, the optimal objective value of Problem (B.4) is $h(N_A)$ (i.e. $\sum_{j=1}^{L} \tau_j \lambda_{A,j}^\dagger = h(N_A)$), where $h(\cdot)$ defined by Equation (B.13), and its optimal solution $\lambda_{A,j}^\dagger$ is equal to $\lambda_j^\star(N_A)$ in Equation (B.12).

Then, by the wage equilibrium, let $N_H^\dagger = \gamma P_H \sum_{j=1}^{L}(\lambda_j^\dagger - \lambda_{A,j}^\dagger)/r$, we want to show the optimal value of Problem (B.5) must be equal to $h(N_A + N_H^\dagger)$ and its optimal solution must be $\lambda_j^\dagger = \lambda_j^\star(N_A + N_H^\dagger)$. Given $N_H^\dagger$, if the overall arrival rate is not maximized (i.e. $\sum_{j=1}^{L} \tau_j \lambda_j^\dagger < h(N_A + N_H^\dagger)$), this means:

$$rN_H^\dagger = \gamma P_H \sum_{j=1}^{L} \tau_j(\lambda_j^\dagger - \lambda_{A,j}^\dagger) < \gamma P_H(h(N_A + N_H^\dagger) - h(N_A))$$

which implies $(h(N_A + N_H^\dagger) - h(N_A))/N_H^\dagger > r/(\gamma P_H)$. By Lemma 15, we know $(h(N_A + N) - h(N_A))/N$ is continuous in $N > 0$. And because $h(N)$ is bounded by $\sum_{j=1}^{L} \tau_j \mu_j$, $(h(N_A + N) - h(N_A))/N$ converges to 0 as $N \to \infty$. Thus, there exists a feasible $\overline{N}_H > N_H^\dagger$ such that $(h(N_A + \overline{N}_H) - h(N_A))/\overline{N}_H = r/(\gamma P_H)$ and $h(N_A + \overline{N}_H) > h(N_A + N_H^\dagger)$, which contradicts the assumption that $N_H^\dagger$ is optimal in Problem (B.5). Therefore, the optimal value of Problem (B.5) must equal $h(N_A + N_H^\dagger)$. And since any solution to Problem (B.5) is feasible in Problem (B.11) given $N_A + N_H^\dagger$, the optimal $\lambda_j^\dagger$ of Problem (B.5) must be equal to $\lambda_j^\star(N_A + N_H^\dagger)$ in Equation (B.12).

Second, given $N_H^\ddagger$ as the optimal equilibrium number of HVs when we fully prioritize HVs, Problem (B.7) is exactly the same with Problem (B.11), so the optimal objective value of Problem (B.7) is $h(N_A + N_H^\ddagger)$, and its optimal solution $\lambda_j^\ddagger$ is equal to $\lambda_j^\star(N_A + N_H^\ddagger)$.

□

In the next lemma, we want to show that if one of the boundary constraints in Problem $(\mathcal{M}')$ is binding (i.e. $\sum_{j=1}^{L} \tau_j \lambda_j = \sum_{j=1}^{L} \tau_j \lambda_j^\dagger$ or $\sum_{j=1}^{L} \tau_j \lambda_j = \sum_{j=1}^{L} \tau_j \lambda_j^\ddagger$), then for any location j, the optimal arrival rate of Problem $(\mathcal{M}')$ must be equal to $\lambda_j^\dagger$ given by Problem

(B.5) or $\lambda_j^\ddagger$ given by Problem (B.7). With a slight abuse of notation, we use $\boldsymbol{\lambda}^\star = \{\lambda_j^\star\}_{j=1}^L$ to denote the optimal solution to Problem $(\mathcal{M}')$.

Lemma 18. If the lower (or upper) boundary constraints in Problem $(\mathcal{M}')$ is binding, the optimal solution must be given by a full prioritization policy. That is, if $\sum_{j=1}^L \tau_j \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\dagger$ (or $\sum_{j=1}^L \tau_j \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\ddagger$), then $\forall j,\ \lambda_j^\star = \lambda_j^\dagger$ (or $\lambda_j^\star = \lambda_j^\ddagger$).

Proof. Proof of Lemma 18. Since Problem $(\mathcal{M}')$ is a convex problem with a strictly concave objective, it must have a unique optimal solution. Thus, if we are able to show that when the lower (or upper) boundary constraint is binding, the objective value achieved by $\boldsymbol{\lambda}^\dagger$ (or $\boldsymbol{\lambda}^\ddagger$) is optimal, then we must have $\forall j,\ \lambda_j^\star = \lambda_j^\dagger$ (or $\forall j,\ \lambda_j^\star = \lambda_j^\ddagger$).

Suppose $\sum_{j=1}^L \tau_j \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\dagger$, and the objective value of Problem $(\mathcal{M}')$ achieved by $\boldsymbol{\lambda}^\star$ is higher than the objective value achieved by $\boldsymbol{\lambda}^\dagger$, then we must have $\sum_{j=1}^L \tau_j \lambda_{A,j}^\star > \sum_{j=1}^L \tau_j \lambda_{A,j}^\dagger$ Because $\sum_{j=1}^L \tau \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\dagger$ and $\sum_{j=1}^L \tau_j \lambda_{A,j}^\star > \sum_{j=1}^L \tau_j \lambda_{A,j}^\dagger$,

$$N_H^\star = \gamma P_H / r \sum_{j=1}^L \tau_j (\lambda_j^\star - \lambda_{A,j}^\star) < \gamma P_H / r \sum_{j=1}^L \tau_j (\lambda_j^\dagger - \lambda_{A,j}^\dagger) = N_H^\dagger$$

However, this means there exists a feasible $N_H^\star < N_H^\dagger$ in Problem (B.5) such that $\sum_{j=1}^L \tau_j \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\dagger = h(N_A + N_H^\dagger)$ since $\sum_{j=1}^L \tau_j \lambda_j^\dagger = h(N_A + N_H^\dagger)$ by Lemma 17. And by definition, $\sum_{j=1}^L \tau_j \lambda_j^\star \leqslant h(N_A + N_H^\star)$, which implies $h(N_A + N_H^\dagger) \leqslant h(N_A + N_H^\star)$ but $N_H^\star < N_H^\dagger$. This contradicts the monotonicity of $h(N)$ shown by Lemma 15. Therefore, if $\sum_{j=1}^L \tau_j \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\dagger$, the objective value achieved by $\boldsymbol{\lambda}^\dagger$ is optimal for Problem $(\mathcal{M}')$.

Similarly, suppose $\sum_{j=1}^L \tau \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\ddagger$, and the objective value achieved by $\boldsymbol{\lambda}^\star$ is higher than the objective value achieved by $\boldsymbol{\lambda}^\ddagger$. We have $\sum_{j=1}^L \tau_j \lambda_{H,j}^\star < \sum_{j=1}^L \tau_j \lambda_{H,j}^\ddagger$, where $\sum_{j=1}^L \tau_j \lambda_{H,j}^\star = \sum_{j=1}^L \tau_j (\lambda_j^\star - \lambda_{A,j}^\star)$. Thus,

$$N_H^\star = \gamma P_H / r \sum_{j=1}^L \tau_j \lambda_{H,j}^\star < \gamma P_H / r \sum_{j=1}^L \tau_j \lambda_{H,j}^\ddagger = N_H^\ddagger$$

However, this means there exists a feasible $N_H^\star < N_H^\ddagger$ in Problem (B.7) such that $\sum_{j=1}^L \tau \lambda_j^\star = \sum_{j=1}^L \tau_j \lambda_j^\ddagger = h(N_A + N_H^\ddagger)$ because $\sum_{j=1}^L \tau_j \lambda_j^\ddagger = h(N_A + N_H^\ddagger)$ by Lemma 17. This contra-

dicts the monotonicity of $h(N)$ shown by Lemma 15. Hence, when one of the boundary constraints is binding, the optimal solution must be given by a full prioritization policy. (i.e. if $\sum_{j=1}^{L} \tau_j \lambda_j^\star = \sum_{j=1}^{L} \tau_j \lambda_j^\dagger$ (or $\sum_{j=1}^{L} \tau_j \lambda_j^\star = \sum_{j=1}^{L} \tau_j \lambda_j^\ddagger$), then for any j, $\lambda_j^\star = \lambda_j^\dagger$ (or $\lambda_j^\star = \lambda_j^\ddagger$).)

□

In the next auxiliary lemma, we want to show that it is implementable to randomize two full prioritization policies and achieve a convex combination of their arrival rates. Before presenting the lemma, let us first define a random-priority policy. Suppose we have two non-idling policies with the same allocation of vehicles, $N_{A,j},\ N_{H,j}$ at each location j, and:

1. π^1: fully prioritizes AVs and achieves the arrival rates $\lambda_{A,j}^{\pi^1}$ and $\lambda_{H,j}^{\pi^1}$ at each location j.
2. π^2: fully prioritizes HVs and achieves the arrival rates $\lambda_{A,j}^{\pi^2}$ and $\lambda_{H,j}^{\pi^2}$ at each location j.

For any $\omega \in [0,1]$, we propose a policy $\bar{\pi}(\theta)$ to randomize π^1 and π^2 and achieve the arrival rates $\omega\lambda_{A,j}^{\pi^1} + (1-\omega)\lambda_{A,j}^{\pi^2}$ and $\omega\lambda_{H,j}^{\pi^1} + (1-\omega)\lambda_{H,j}^{\pi^2}$ for AVs and HVs respectively.

Definition 5 (Definition of random-priority policy $\bar{\pi}(\theta)$)**.** Keep the allocation of vehicles $N_{A,j},\ N_{H,j}$ the same with those in π^1 and π^2. At each location, there is a non-idling priority queue and a non-idling non-priority queue. The vehicles in the priority queue are dispatched before those in the non-priority queue. Given $\theta \in [0,1]$, an incoming AV is allocated in the priority queue with probability θ and in the non-priority queue with probability $1-\theta$. Accordingly, an incoming HV is allocated in the priority queue with probability $1-\theta$ and in the non-priority queue with probability θ.

Let $\bar{\lambda}_{A,j},\ \bar{\lambda}_{H,j}$ denote the arrival rates of AVs and HVs induced by $\bar{\pi}(\theta)$ at location j. The following lemma shows that by choosing $\theta \in [0,1]$ and implementing $\bar{\pi}(\theta)$, we can randomize π^1 and π^2 to obtain the arrival rates $\omega\lambda_{A,j}^{\pi^1} + (1-\omega)\lambda_{A,j}^{\pi^2}$ and $\omega\lambda_{H,j}^{\pi^1} + (1-\omega)\lambda_{H,j}^{\pi^2}$ for AVs and HVs respectively.

Lemma 19 (Randomization of two prioritization policies)**.** For any $\omega \in [0,1]$, there exists $\theta \in [0,1]$ such that we are able to achieve $\bar{\lambda}_{A,j} = \omega\lambda^{\pi^1}_{A,j} + (1-\omega)\lambda^{\pi^2}_{A,j}$ and $\bar{\lambda}_{H,j} = \omega\lambda^{\pi^1}_{H,j} + (1-\omega)\lambda^{\pi^2}_{H,j}$ by implementing $\bar{\pi}(\theta)$.

Proof. Proof of Lemma 19 Since there are two variables $\bar{\lambda}_{A,j}$ and $\bar{\lambda}_{H,j}$ for each location j, we first want to reduce the variable space into $\bar{\lambda}_{A,j}$ only. That is, we want to show that if $\bar{\lambda}_{A,j}$ is determined and satisfies Little's law, then $\bar{\lambda}_{H,j}$ is also determined and satisfies Little's law. After that, we want to show the solution to $\bar{\lambda}_{A,j}$ is continuous with respect to θ and apply the intermediate value theorem to complete the proof.

Let λ^p_j and λ^{np}_j denote the arrival rates of vehicles in the priority queue and non-priority queue at each location j. By Definition 5, $\lambda^p_j = \theta\bar{\lambda}_{A,j}+(1-\theta)\bar{\lambda}_{H,j}$ and $\lambda^{np}_j = (1-\theta)\bar{\lambda}_{A,j}+\theta\bar{\lambda}_{H,j}$.

Because the arrival of requests is independent of vehicles, the mean residual service time is equal to $1/\mu$. Then, by Haviv (2013), pp. 73-74, the mean waiting time of the vehicles in the priority queue of location j equals

$$W^{\bar{\pi},p}_j = \frac{1/\mu_j}{1-\lambda^p_j/\mu_j} = \frac{1/\mu_j}{1-\lambda^p_j/\mu_j} = \frac{1}{\mu_j-\lambda^p_j}$$

And the mean waiting time of the vehicles in the non-priority queue of location j equals

$$W^{\bar{\pi},np}_j = \frac{1/\mu_j}{(1-\lambda^p_j/\mu_j)(1-(\lambda^p_j+\lambda^{np}_j)/\mu_j)} = \frac{\mu_j}{(\mu_j-\lambda^p_j)(\mu_j-(\lambda^p_j+\lambda^{np}_j))}$$

Thus, the unconditional mean waiting time of AVs and HVs are:

$$W^{\bar{\pi}}_{A,j} = \theta W^{\bar{\pi},p}_j + (1-\theta)W^{\bar{\pi},np}_j = \frac{\mu_j-\theta(\lambda^p_j+\lambda^{np}_j)}{(\mu_j-\lambda^p_j)(\mu_j-(\lambda^p_j+\lambda^{np}_j))}$$
$$W^{\bar{\pi}}_{H,j} = (1-\theta)W^{\bar{\pi},p}_j + (\theta)W^{\bar{\pi},np}_j = \frac{\mu_j-(1-\theta)(\lambda^p_j+\lambda^{np}_j)}{(\mu_j-\lambda^p_j)(\mu_j-(\lambda^p_j+\lambda^{np}_j))}$$

Substitute $\lambda^p_j = \theta\bar{\lambda}_{A,j}+(1-\theta)\bar{\lambda}_{H,j}$ and $\lambda^{np}_j = (1-\theta)\bar{\lambda}_{A,j}+(\theta)\bar{\lambda}_{H,j}$ into the above equations:

$$W^{\bar{\pi}}_{A,j} = \frac{\mu_j-\theta(\bar{\lambda}_{A,j}+\bar{\lambda}_{H,j})}{(\mu_j-(\theta\bar{\lambda}_{A,j}+(1-\theta)\bar{\lambda}_{H,j}))(\mu_j-(\bar{\lambda}_{A,j}+\bar{\lambda}_{H,j}))}$$
$$W^{\bar{\pi}}_{H,j} = \frac{\mu_j-(1-\theta)(\bar{\lambda}_{A,j}+\bar{\lambda}_{H,j})}{(\mu_j-(\theta\bar{\lambda}_{A,j}+(1-\theta)\bar{\lambda}_{H,j}))(\mu_j-(\bar{\lambda}_{A,j}+\bar{\lambda}_{H,j}))}$$

In addition, since the policies are non-idling and the number of vehicles at each location is fixed, the total arrival rate of vehicles must be the same. By Little's law, we must have a constant $\lambda_j \in [0, \mu_j)$ such that $(\tau_j + \frac{1}{\mu_j - \lambda_j})\lambda_j = N_{A,j} + N_{H,j} = N_j$, and $\lambda^{\pi^1}_{A,j} + \lambda^{\pi^1}_{H,j} = \lambda^{\pi^2}_{A,j} + \lambda^{\pi^2}_{H,j} = \lambda^p_j + \lambda^{np}_j = \bar{\lambda}_{A,j} + \bar{\lambda}_{H,j} = \lambda_j$. Then, we can rewrite the mean waiting time as:

$$W^{\bar{\pi}}_{A,j} = \frac{\mu_j - \theta\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))(\mu_j - \lambda_j)}$$
$$W^{\bar{\pi}}_{H,j} = \frac{\mu_j - (1-\theta)\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))(\mu_j - \lambda_j)}$$

By Little's law,

$$(\tau_j + \frac{\mu_j - \theta\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))(\mu_j - \lambda_j)})\bar{\lambda}_{A,j} = N_{A,j} \tag{B.23a}$$
$$(\tau_j + \frac{\mu_j - (1-\theta)\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))(\mu_j - \lambda_j)})(\lambda_j - \bar{\lambda}_{A,j}) = N_j - N_{A,j} \tag{B.23b}$$

In Equation (B.23b), if we substitute N_j with $(\tau_j + \frac{1}{\mu_j - \lambda_j})\lambda_j$ and $N_{A,j}$ with Equation (B.23a), we can see Equation (B.23a) and Equation (B.23b) are dependent:

$$\begin{aligned}
N_j - N_{A,j} &= (\tau_j + \frac{1}{\mu_j - \lambda_j})\lambda_j - (\tau_j + \frac{\mu_j - \theta\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))(\mu_j - \lambda_j)})\bar{\lambda}_{A,j} \\
&= \tau_j(\lambda_j - \bar{\lambda}_{A,j}) + \frac{\lambda_j(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j)) - (\mu_j - \theta\lambda_j)\bar{\lambda}_{A,j}}{(\mu_j - \lambda_j)(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))} \\
&= \tau_j(\lambda_j - \bar{\lambda}_{A,j}) + \frac{\mu_j(\lambda_j - \bar{\lambda}_{A,j}) - \lambda_j((1-\theta)\lambda_j + (\theta-1)\bar{\lambda}_{A,j}))}{(\mu_j - \lambda_j)(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))} \\
&= \tau_j(\lambda_j - \bar{\lambda}_{A,j}) + \frac{(\mu_j - (1-\theta)\lambda_j)(\lambda_j - \bar{\lambda}_{A,j})}{(\mu_j - \lambda_j)(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))} \\
&= (\tau_j + \frac{\mu_j - (1-\theta)\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))(\mu_j - \lambda_j)})(\lambda_j - \bar{\lambda}_{A,j})
\end{aligned}$$

Thus, we only need $\bar{\lambda}_{A,j}$ to be a solution to Equation (B.23a), and it must be a solution to Equation (B.23b) as well. In other words, if $\bar{\lambda}_{A,j}$ is determined and satisfies Little's law, then $\bar{\lambda}_{H,j}$ equals $\lambda_j - \bar{\lambda}_{A,j}$ and also satisfies Little's law.

Second, we want to show that $\bar{\lambda}_{A,j}$ of Equation (B.23a) is continuous with respect to θ so that we can apply the intermediate value theorem to demonstrate our main argument. For

any $\theta \in [0,1]$, let $\bar{\lambda}_{A,j}(\theta)$ denote a solution to Equation (B.23a). Notice that when $\theta = 0$, $\bar{\lambda}_{A,j}(0)$ is a solution to

$$(\tau_j + \frac{\mu_j}{\mu_j - \bar{\lambda}_{A,j}})\bar{\lambda}_{A,j} = N_{A,j}$$

Thus, $\bar{\lambda}_{A,j}(0) = \lambda^{\pi^1}_{A,j}$. Similarly, $\bar{\lambda}_{A,j}(1) = \lambda^{\pi^2}_{A,j}$. Because $\omega\lambda^{\pi^1}_{A,j} + (1-\omega)\lambda^{\pi^2}_{A,j} \in [\lambda^{\pi^1}_{A,j}, \lambda^{\pi^2}_{A,j}]$ for some $\omega \in [0,1]$, if we are able to show $\bar{\lambda}_{A,j}(\theta)$ is continuous, then there must be a $\theta \in [0,1]$ such that $\bar{\lambda}_{A,j}(\theta) = \omega\lambda^{\pi^1}_{A,j} + (1-\omega)\lambda^{\pi^2}_{A,j}$ by the intermediate value theorem.

Define

$$F(\theta, \bar{\lambda}_{A,j}) = (\tau_j + \frac{\mu_j - \theta\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))(\mu_j - \lambda_j)})\bar{\lambda}_{A,j} - N_{A,j}$$

on the domain $[0,1] \times [0, \lambda_j]$. Because $\lambda_j < \mu_j$, $F(\theta, \bar{\lambda}_{A,j})$ is continuously differentiable on the domain. And for each fixed θ, $F(\theta, 0) = -N_{A,j} \leqslant 0$, $F(\theta, \lambda_j) = (\tau_j + \frac{1}{\mu_j - \lambda_j})\lambda_j - N_{A,j} = N_{H,j} \geqslant 0$. Also, by taking the gradient, $\frac{\partial F(\theta, \bar{\lambda}_{A,j})}{\partial \bar{\lambda}_{A,j}} = \tau_j + \frac{\mu_j - (1-\theta)\lambda_j}{(\mu_j - ((2\theta-1)\bar{\lambda}_{A,j} + (1-\theta)\lambda_j))^2} \cdot \frac{\mu_j - \theta\lambda_j}{\mu_j - \lambda_j} > 0$. Consequently, for each fixed $\theta \in [0,1]$, there is a unique solution $\bar{\lambda}_{A,j}(\theta) \in [0, \lambda_j)$ such that $F(\theta, \bar{\lambda}_{A,j}(\theta)) = 0$. Because of the implicit function theorem and $\frac{\partial F(\theta, \bar{\lambda}_{A,j}(\theta))}{\partial \bar{\lambda}_{A,j}} > 0$, $\bar{\lambda}_{A,j}(\theta)$ is a continuous function of θ.

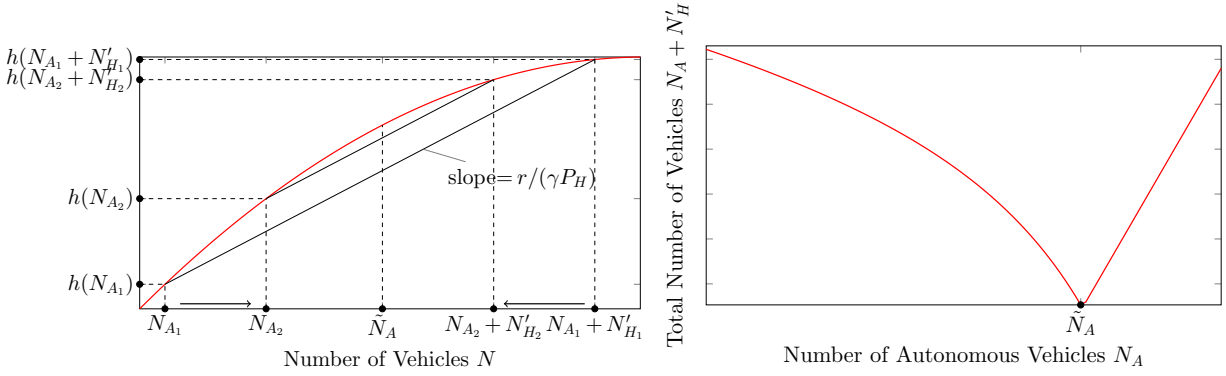
Hence, by the intermediate value theorem, there must be a $\theta \in [0,1]$ such that $\bar{\lambda}_{A,j}(\theta) = \omega\lambda^{\pi^1}_{A,j} + (1-\omega)\lambda^{\pi^2}_{A,j}$. In addition, because $\lambda^{\pi^1}_{A,j} + \lambda^{\pi^1}_{H,j} = \lambda^{\pi^2}_{A,j} + \lambda^{\pi^2}_{H,j} = \bar{\lambda}_{A,j} + \bar{\lambda}_{H,j} = \lambda_j$, we must have $\bar{\lambda}_{H,j}(\theta) = \lambda_j - \bar{\lambda}_{A,j}(\theta) = \omega\lambda^{\pi^1}_{H,j} + (1-\omega)\lambda^{\pi^2}_{H,j}$.

□

Given the above auxiliary lemmas, now we can continue to complete the proof of Proposition 7.

B.2.3 Supporting Material of the Main Results in Section 3.4 and Section 3.5.

In this section, we prove several auxiliary lemmas for our main results in Section 3.4 and Section 3.5. We first discuss the impact of fully prioritizing AVs on the total number of vehicles. To help readers better understand it, we visualize the reasoning in Figure B.2.



(a) (b)

Figure B.2: Visualization of Lemma 20. The left plot is an example of $h(N)$, and the right plot shows the change in the total number of vehicles with respect to N_A.

Given N_A, when we fully prioritize AVs, the optimal arrival rate of AVs is equal to $h(N_A)$ and the optimal arrival rate of HVs is equal to $h(N_A + N_H^\dagger) - h(N_A)$. With the wage equilibrium, we have $\frac{h(N_A+N_H^\dagger)-h(N_A)}{N_H^\dagger} = r/(\gamma P_H)$, so the slope of the line connecting $(N_A, h(N_A))$ and $(N_A + N_H^\dagger, h(N_A + N_H^\dagger))$ must be equal to $r/(\gamma P_H)$. Thus, as shown in the left plot of Figure B.2, when we increase the number of AVs from N_{A_1} to N_{A_2}, due to the strict concavity of $h(N)$, the line connecting $(N_{A_2}, h(N_{A_2}))$ and $(N_{A_2} + N_{H_2}^\dagger, h(N_{A_2} + N_{H_2}^\dagger))$ must be above the line connecting $(N_{A_1}, h(N_{A_1}))$ and $(N_{A_1} + N_{H_1}^\dagger, h(N_{A_1} + N_{H_1}^\dagger))$. Accordingly, the total number of vehicles will decrease from $N_{A_1} + N_{H_1}^\dagger$ to $N_{A_2} + N_{H_2}^\dagger$ as shown in the left plot.

In addition, let $\widetilde{N}_A > 0$ such that $h'(\widetilde{N}_A) = r/(\gamma P_H)$. Because $h(\cdot)$ is strictly concave, for any $N_A \geqslant \widetilde{N}_A$, $h'(N_A) < r/(\gamma P_H)$, meaning that it is impossible to have a positive N_H such that the line connecting $(N_A, h(N_A))$ and $(N_A + N_H, h(N_A + N_H))$ has a slope of $r/(\gamma P_H)$. Thus, HVs completely disappear after the number of AVs reaches $\widetilde{N}_A$, and an increase of N_A will lead to a higher total number of vehicles.

Now let us verify the idea in a formal way. The following lemma shows the property of the optimal solution to N_H when we fully prioritize AVs. To emphasize the relationship

with N_A and γ, we use $N_H^\dagger(N_A, \gamma)$ instead of $N_H^\dagger$ to denote the optimal number of HVs in Problem (B.5) given N_A and γ.

Lemma 20 (Optimal N_H when we fully prioritize AVs)**.** When we fully prioritize AVs, the optimal arrival rate of AVs is $h(N_A)$. And $N_H^\dagger(N_A, \gamma)$ has the following properties with respect to N_A and γ:

(a) $N_H^\dagger(N_A, \gamma)$ is positive and unique if and only if $h'(N_A) > r/(\gamma P_H)$. In addition, if $N_H^\dagger(N_A, \gamma) > 0$, then $h'(N_A + N_H^\dagger(N_A, \gamma)) < r/(\gamma P_H)$.

(b) Let $\tilde{\gamma} > 0$ and $\widetilde{N}_A > 0$ such that $h'(\widetilde{N}_A) = r/(\tilde{\gamma} P_H)$. Then, given any $\gamma > \tilde{\gamma}$, $\frac{\partial N_H^\dagger(N_A,\gamma)}{\partial N_A} < -1$ if $N_A \in [0, \widetilde{N}_A)$ and $N_H^\dagger(N_A, \gamma) = 0$ if $N_A \in [\widetilde{N}_A, \infty)$. In addition, given any $N_A \in (0, \widetilde{N}_A)$, $\frac{\partial N_H^\dagger(N_A,\gamma)}{\partial \gamma} > 0$ if $\gamma > \tilde{\gamma}$, and $N_H^\dagger(N_A, \gamma) \to \infty$ as $\gamma \to \infty$.

Proof. Proof of Lemma 20.

(a) To see this, we first notice that Problem (B.4) is equivalent to Problem (B.11) with $N = N_A$, so the results in the Lemma 14 and Lemma 15 are applicable, and the optimal value of Problem (B.4) can be written as $h(N_A)$, where $h(\cdot)$ is given by Equation (B.13). Also, Problem (B.5) can be rewritten as:

$$\begin{aligned} \max_{N_H \geqslant 0} \quad & \left(\max_{\lambda_j \in [0, \mu_j)} \sum_{j=1}^{L} \tau_j \lambda_j \quad \text{s.t.} \quad \sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_j}) \lambda_j = N_A + N_H \right) \\ \text{s.t.} \quad & r N_H = \gamma P_H \sum_{j=1}^{L} \tau_j (\lambda_j - \lambda_{A,j}^\dagger) \end{aligned}$$

We can see that Problem (B.11) is actually a subproblem of Problem (B.5) with $N = N_A + N_H$. Therefore, by Lemma 17, Problem (B.5) is equivalent to:

$$\begin{aligned} \max_{N_H \geqslant 0} \quad & h(N_A + N_H) \\ \text{s.t.} \quad & r N_H = \gamma P_H [h(N_A + N_H) - h(N_A)] \end{aligned} \tag{B.24}$$

It is easy to see that $N_H = 0$ is always feasible in Problem (B.24), but any feasible positive N_H will produce a higher objective value, since $h(N)$ strictly increases by

Lemma 15. Now we want to obtain the sufficient and necessary conditions such that there exists a unique non-trivial optimal N_H.

- $N^\dagger(N_A,\gamma) > 0 \implies h'(N_A) > r/(\gamma P_H)$ and $h'(N_A + N_H^\dagger(N_A,\gamma)) < r/(\gamma P_H)$.

  By the mean value theorem, if there exists $N_H > 0$ such that $rN_H = \gamma P_H[h(N_A + N_H) - h(N_A)]$, we must have $\exists n \in (N_A, N_A + N_H)$ such that

  $$h'(n) = \frac{h(N_A + N_H) - h(N_A)}{N_H} = \frac{r}{\gamma P_H}$$

  However, because $h(N)$ is strictly concave by Lemma 15, if $h'(N_A) \leqslant r/(\gamma P_H)$, then $\forall n > N_A,\ h'(n) < r/(\gamma P_H)$, which means it is impossible to have a $N_H > 0$ such that $rN_H = \gamma P_H[h(N_A+N_H)-h(N_A)]$. Similarly, if $h'(N_A+N_H) \geqslant r/(\gamma P_H)$, then $\forall n < N_A + N_H,\ h'(n) > r/(\gamma P_H)$, which means it is impossible to have a $N_H > 0$ such that $rN_H = \gamma P_H[h(N_A + N_H) - h(N_A)]$. Thus, $N^\dagger(N_A,\gamma) > 0 \implies h'(N_A) > r/(\gamma P_H)$ and $h'(N_A + N_H^\dagger(N_A,\gamma)) < r/(\gamma P_H)$.

- $h'(N_A) > r/(\gamma P_H) \implies N^\dagger(N_A,\gamma) > 0$

  Suppose $h'(N_A) > r/(\gamma P_H)$, then $\lim_{N_H\to 0} \frac{h(N_A+N_H)-h(N_A)}{N_H} = h'(N_A) > r/(\gamma P_H)$. Thus, it is sufficient to show $\frac{h(N_A+N_H)-h(N_A)}{N_H}$ strictly decreases in N_H and converges to 0 as $N \to \infty$. Let $n_1 > n_2 > 0$, because $h(N)$ is strictly concave,

  $$\begin{aligned} & h(N_A + n_2) > \frac{n_1 - n_2}{n_1} h(N_A) + \frac{n_2}{n_1} h(N_A + n_1) \\ \implies & h(N_A + n_2) - h(N_A) > \frac{n_2}{n_1}(h(N_A + n_1) - h(N_A)) \\ \implies & \frac{h(N_A + n_2) - h(N_A)}{n_2} > \frac{h(N_A + n_1) - h(N_A)}{n_1} \end{aligned} \tag{B.25}$$

  And by Equation (B.17), $\lim_{N\to\infty} h(N) = \sum_{j=1}^{L} \tau_j \mu_j$, so $\lim_{N_H\to\infty} \frac{h(N_A+N_H)-h(N_A)}{N_H} = 0$. Therefore, by the intermediate value theorem, there exists a unique $N_H^\dagger(N_A,\gamma) > 0$ such that

  $$\frac{h(N_A + N_H^\dagger(N_A,\gamma)) - h(N_A)}{N_H^\dagger(N_A,\gamma)} = \frac{r}{\gamma P_H}$$

(b) Now, suppose $\tilde{\gamma} > 0$ and $\widetilde{N}_A > 0$ such that $h'(\widetilde{N}_A) = r/(\tilde{\gamma} P_H)$. For any $N_A \geqslant \widetilde{N}_A$, because $h(N)$ is stictly concave, $h'(N_A) \leqslant r/(\tilde{\gamma} P_H)$, so $N_H^\dagger(N_A, \gamma) = 0$ and $N_A + N_H^\dagger(N_A, \gamma) = N_A$.

The remaining step is to show that $N_A + N_H^\dagger(N_A, \gamma)$ is strictly decreasing in $N_A \in [0, \widetilde{N}_A)$ and strictly increasing in $\gamma \in (\tilde{\gamma}, \infty)$.

First, given $\gamma > \tilde{\gamma}$, let $N_A \in [0, \widetilde{N}_A)$, by the wage equilibrium:

$$
\begin{aligned}
& h(N_A + N_H^\dagger(N_A, \gamma)) - h(N_A) - r/(\gamma P_H) N_H^\dagger(N_A, \gamma) = 0 \\
\Longrightarrow\ & h'(N_A + N_H^\dagger(N_A, \gamma))(1 + \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial N_A}) - h'(N_A) - r/(\gamma P_H) \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial N_A} = 0 \\
& \text{Because } h'(N_A) > r/(\gamma \tilde{P}_H) > r/(\gamma P_H) \text{ and } h'(N_A + N_H^\dagger(N_A, \gamma)) < r/(\gamma P_H), \\
\Longrightarrow\ & \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial N_A} = \frac{h'(N_A) - h'(N_A + N_H^\dagger(N_A, \gamma))}{h'(N_A + N_H^\dagger(N_A, \gamma)) - r/(\gamma P_H)} < -1 \\
\Longrightarrow\ & \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial N_A} < -1
\end{aligned}
$$

Second, given $N_A \in (0, \widetilde{N}_A)$, let $\gamma > \tilde{\gamma}$, by the wage equilibrium:

$$
\begin{aligned}
& h(N_A + N_H^\dagger(N_A, \gamma)) - h(N_A) - r/(\gamma P_H) N_H^\dagger(N_A, \gamma) = 0 \\
\Longrightarrow\ & h'(N_A + N_H^\dagger(N_A, \gamma)) \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial \gamma} \\
& - [-r/(\gamma^2 P_H) N_H^\dagger(N_A, \gamma) + r/(\gamma P_H) \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial \gamma}] = 0 \\
& \text{Because } h'(N_A + N_H^\dagger(N_A, \gamma)) < r/(\gamma P_H), \\
\Longrightarrow\ & \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial \gamma} = \frac{r/(\gamma^2 P_H) N_H^\dagger(N_A, \gamma)}{r/(\gamma P_H) - h'(N_A + N_H^\dagger(N_A, \gamma))} > 0 \\
\Longrightarrow\ & \frac{\partial N_H^\dagger(N_A, \gamma)}{\partial \gamma} > 0
\end{aligned}
$$

Additionally, in part (a), we have seen that $\frac{h(N_A+N_H)-h(N_A)}{N_H}$ strictly decreases in N_H and $\frac{h(N_A+N_H)-h(N_A)}{N_H} \to 0$ as $N_H \to \infty$. Thus, with $\frac{h(N_A+N_H^\dagger(N_A,\gamma))-h(N_A)}{N_H^\dagger(N_A,\gamma)} = r/(\gamma P_H)$, the increase of $N_H^\dagger(N_A, \gamma)$ must be unbounded as $\gamma \to \infty$. That is, $N_H^\dagger(N_A, \gamma) \to \infty$ as $\gamma \to \infty$.

□

The following lemma shows the service levels when AVs are fully prioritized, and N_A reaches the threshold such that all HVs leave the market.

Lemma 21. When AVs are fully prioritized, let $\widetilde{N}_A > 0$ as the treshold such that $N_H^\dagger(N_A, \gamma) = 0$ for any $N_A \geqslant \widetilde{N}_A$, then

$$\lambda_j^\star(\widetilde{N}_A) = [\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\sqrt{\frac{r}{\gamma P_H - r}}]\mathbf{1}_{\mu_j\tau_j \geqslant r/(\gamma P_H - r)} \quad \forall j \in \{1, \dots, L\} \tag{B.26}$$

Proof. Proof of Lemma 21. By Lemma 20, $h'(\widetilde{N}_A) = r/(\gamma P_H)$. Using Equation (B.21),

$$\begin{aligned} &\frac{1}{2}(1 - \frac{d_k(\widetilde{N}_A)}{\sqrt{\Delta_k(\widetilde{N}_A)}}) = r/(\gamma P_H) \\ \implies &\frac{d_k(\widetilde{N}_A)}{\sqrt{\Delta_k(\widetilde{N}_A)}} = 1 - \frac{2r}{\gamma P_H} \end{aligned} \tag{B.27}$$

And with Equation (B.12),

$$\lambda_j^\star(\widetilde{N}_A) = [\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\frac{-d_k(\widetilde{N}_A) + \sqrt{\Delta_k(\widetilde{N}_A)}}{2(\sum_{j=1}^{k(\widetilde{N}_A)}\sqrt{\mu_j\tau_j})}]\mathbf{1}_{j \leqslant k(\widetilde{N}_A)}$$

where $d_k(\widetilde{N}_A) = \widetilde{N}_A + k(\widetilde{N}_A) - \sum_{j=1}^{k(\widetilde{N}_A)}\mu_j\tau_j$, $\Delta_k(\widetilde{N}_A) = d_k(\widetilde{N}_A)^2 + 4(\sum_{j=1}^{k(\widetilde{N}_A)}\sqrt{\mu_j\tau_j})^2$. Using the definitions of $d_k(\widetilde{N}_A)$ and $\Delta_k(\widetilde{N}_A)$,

$$\begin{aligned} \frac{-d_k(\widetilde{N}_A) + \sqrt{\Delta_k(\widetilde{N}_A)}}{2(\sum_{j=1}^{k(\widetilde{N}_A)}\sqrt{\mu_j\tau_j})} &= \frac{-d_k(\widetilde{N}_A) + \sqrt{\Delta_k(\widetilde{N}_A)}}{\sqrt{\Delta_k(\widetilde{N}_A) - d_k(\widetilde{N}_A)^2}} \\ &= \sqrt{\frac{\sqrt{\Delta_k(\widetilde{N}_A)} - d_k(\widetilde{N}_A)}{\sqrt{\Delta_k(\widetilde{N}_A)} + d_k(\widetilde{N}_A)}} \\ &= \sqrt{\frac{1 - d_k(\widetilde{N}_A)/\sqrt{\Delta_k(\widetilde{N}_A)}}{1 + d_k(\widetilde{N}_A)/\sqrt{\Delta_k(\widetilde{N}_A)}}} \\ &= \sqrt{\frac{r}{\gamma P_H - r}} \quad \text{substitute } d_k(\widetilde{N}_A)/\sqrt{\Delta_k(\widetilde{N}_A)} \text{ with Equation (B.27)} \end{aligned}$$

Therefore,

$$\lambda_j^\star(\widetilde{N}_A) = [\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\sqrt{\frac{r}{\gamma P_H - r}}]\mathbf{1}_{\mu_j\tau_j \geqslant r/(\gamma p - r)} \quad \forall j \in \{1, \ldots, L\}$$

Notice that we are able to replace $\mathbf{1}_{j \leqslant k(\widetilde{N}_A)}$ with $\mathbf{1}_{\mu_j\tau_j \geqslant r/(\gamma P_H - r)}$ because of the definition of $k(N)$ in Equation (B.18) and $\frac{-d_k(\widetilde{N}_A)+\sqrt{\Delta_k(\widetilde{N}_A)}}{2(\sum_{j=1}^{k(\widetilde{N}_A)}\sqrt{\mu_j\tau_j})} = \sqrt{\frac{r}{\gamma P_H - r}}$ above.

□

In the next lemma, we discuss the properties of the service levels induced by Equation (B.12). Let $\rho_j^\star(N)$ denote the service level at location j implied by Equation (B.12):

$$\rho_j^\star(N) = \frac{\lambda_j^\star(N)}{\mu_j} = [1 - \sqrt{\frac{1}{\mu_j\tau_j}}\frac{-d_k(N) + \sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}]\mathbf{1}_{j \leqslant k(N)} \tag{B.28}$$

where $k(N),\ d_k(N)$ and $\Delta_k(N)$ are defined in Lemma 14.

For the service levels induced by Equation (B.12), Lemma 22 illustrates that high-demand areas have a higher service level than low-demand areas. Also, for locations where vehicles exist, the change in service level in a low-demand area with respect to the number of vehicles is larger than the change in a high-demand area. Recall that we assumed $\forall j \in \{1, \ldots, L-1\},\ \mu_j\tau_j \geqslant \mu_{j+1}\tau_{j+1}$.

Lemma 22. $\forall N \geqslant 0,\ \rho_j^\star(N) \geqslant \rho_{j+1}^\star(N)$, and if $j+1 \leqslant k(N)$, $\frac{\partial\rho_j^\star(N)}{\partial N} \leqslant \frac{\partial\rho_{j+1}^\star(N)}{\partial N}$.

Proof. Proof of Lemma 22. Recall that Lemma 14 showed that there exists a non-decreasing sequence $\{\overline{N}_j\}_{j=1}^L$ such that $\forall N \leqslant \overline{N}_j,\ \lambda_j^\star(N) = 0$, and $\forall N > \overline{N}_j,\ \lambda_j^\star(N) > 0$. In addition, we defined $k(N) = \max\{k|N > \overline{N}_k\}$, then:

$$\begin{aligned}
&\rho_j^\star(N) - \rho_{j+1}^\star(N) \\
&= \mathbf{1}_{j \leqslant k(N)} - \mathbf{1}_{j+1 \leqslant k(N)} + [\sqrt{\frac{1}{\mu_{j+1}\tau_{j+1}}}\mathbf{1}_{j+1 \leqslant k(N)} - \sqrt{\frac{1}{\mu_j\tau_j}}\mathbf{1}_{j \leqslant k(N)}]\frac{-d_k(N) + \sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}
\end{aligned}$$

Now there are three cases:

1. If $j > k(N)$, $\rho_j^\star(N) = \rho_{j+1}^\star(N) = 0$

2. If $j+1 > k(N)$ but $j \leqslant k(N)$, $\rho_j^\star(N) - \rho_{j+1}^\star(N) = \rho_j^\star(N) \geqslant 0$.

3. If $j+1 \leqslant k(N)$,

$$\rho_j^\star(N) - \rho_{j+1}^\star(N) = [\sqrt{\frac{1}{\mu_{j+1}\tau_{j+1}}} - \sqrt{\frac{1}{\mu_j\tau_j}}]\frac{-d_k(N) + \sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})} \geqslant 0$$

   The non-negativity is because $\mu_j\tau_j \geqslant \mu_{j+1}\tau_{j+1}$.

Thus, we have $\rho_j^\star(N) \geqslant \rho_{j+1}^\star(N)$.

Second, by taking the derivative of $\rho_j^\star(N)$,

$$\frac{\partial\rho_j^\star(N)}{\partial N} = (\sqrt{\frac{1}{\mu_j\tau_j}} \cdot \frac{1}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})})(1 - \frac{d_k(N)}{\sqrt{\Delta_k(N)}})\mathbf{1}_{j\leqslant k(N)}$$

Note that there is a point of N where $\rho_j^\star(N)$ is not differentiable (i.e. when $j = k(N)$), but this does not affect our analysis, or we can consider the right-hand derivative at this point.

Recall that, in Equation (B.16), we have seen $(1 - \frac{d_k(N)}{\sqrt{\Delta_k(N)}}) > 0$. Again, there are three cases:

1. If $j > k(N)$, $\frac{\partial\rho_j^\star(N)}{\partial N} = \frac{\partial\rho_{j+1}^\star(N)}{\partial N} = 0$

2. If $j+1 > k(N)$ but $j \leqslant k(N)$, $\frac{\partial\rho_j^\star(N)}{\partial N} - \frac{\partial\rho_{j+1}^\star(N)}{\partial N} = \frac{\partial\rho_j^\star(N)}{\partial N} \geqslant 0$.

3. If $j+1 \leqslant k(N)$,

$$\begin{aligned}
&\frac{\partial\rho_j^\star(N)}{\partial N} - \frac{\partial\rho_{j+1}^\star(N)}{\partial N} \\
&= (\sqrt{\frac{1}{\mu_j\tau_j}} - \sqrt{\frac{1}{\mu_{j+1}\tau_{j+1}}}) \cdot \frac{1}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}(1 - \frac{d_k(N)}{\sqrt{\Delta_k(N)}}) \\
&\leqslant 0
\end{aligned}$$

   The non-positivity is because $\mu_j\tau_j \geqslant \mu_{j+1}\tau_{j+1}$.

Thus, when $j+1 \leqslant k(N)$, $\frac{\partial \rho_j^\star(N)}{\partial N} \leqslant \frac{\partial \rho_{j+1}^\star(N)}{\partial N}$.

□

In the next lemma, we discuss the number of vehicles induced by Equation (B.12). Let $N_j^\star(N)$ denote the number of vehicles at location j implied by Equation (B.12). By Lemma 2 and Little's law, it must satisfy:

$$N_j^\star(N) = (\tau_j + \frac{1}{\mu_j - \lambda_j^\star(N)})\lambda_j^\star(N) \tag{B.29}$$

For the number of vehicles induced by Equation (B.12), Lemma 23 illustrates that high-demand areas can acquire more vehicles than low-demand areas.

Lemma 23. $\forall j \in \{2, \ldots, L\}, \frac{\partial N_{j-1}^\star(N)}{\partial N} \geqslant \frac{\partial N_j^\star(N)}{\partial N} \geqslant 0.$

Proof. Proof of Lemma 23. Recall we assumed that for any $j \in \{2, \ldots, L\}$, $\mu_{j-1}\tau_{j-1} \geqslant \mu_j\tau_j$. And recall that Lemma 14 showed that there exists a non-decreasing sequence $\{\overline{N}_j\}_{j=1}^L$ such that $\forall N \leqslant \overline{N}_j,\ \lambda_j^\star(N) = 0$, and $\forall N > \overline{N}_j,\ \lambda_j^\star(N) > 0$. In addition, we defined $k(N) = \max\{k | N > \overline{N}_k\}$.

By Equation (B.12):

$$\lambda_j^\star(N) = [\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\frac{-d_k(N) + \sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}]\mathbf{1}_{j \leqslant k(N)}$$

where $d_k(N) = N + k(N) - \sum_{j=1}^{k(N)} \mu_j\tau_j$ and $\Delta_k(N) = d_k(N)^2 + 4(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})^2$.

Clearly, if $j > k(N)$, $\lambda_j^\star(N) = 0$ and $\frac{\partial N_j^\star(N)}{\partial N} = 0$, so we have $\frac{\partial N_{j-1}^\star(N)}{\partial N} \geqslant \frac{\partial N_j^\star(N)}{\partial N} = 0$.

Suppose $j \leqslant k(N)$, let $c_k(N) = \frac{-d_k(N)+\sqrt{\Delta_k(N)}}{2(\sum_{j=1}^{k(N)}\sqrt{\mu_j\tau_j})}$, then

$$\lambda_j^\star(N) = \mu_j - c_k(N)\sqrt{\frac{\mu_j}{\tau_j}}$$

Substitute the above equation into Equation (B.29):

$$\begin{aligned} N_j^\star(N) &= (\tau_j + \frac{1}{\mu_j - \lambda_j^\star(N)})\lambda_j^\star(N) \\ &= (\tau_j + \frac{1}{c_k(N)}\sqrt{\frac{\tau_j}{\mu_j}})(\mu_j - c_k(N)\sqrt{\frac{\mu_j}{\tau_j}}) \end{aligned}$$

Taking the derivative of $N_j^\star(N)$ with respect to N:

$$\frac{\partial N_j^\star(N)}{\partial N} = \sqrt{\mu_j \tau_j}(c_k(N)^{-2} + 1)(-\frac{\partial c_k(N)}{\partial N})$$

Note that there is a point of N where $N_j^\star(N)$ is not differentiable (i.e. when $j = k(N)$), but this does not affect our analysis, or we can consider the right-hand derivative at this point.

We can get $c_k(N)^{-2} > 0$, and by Equation (B.16), $\frac{\partial c_k(N)}{\partial N} < 0$. Because $\sqrt{\mu_{j-1}\tau_{j-1}} \geqslant \sqrt{\mu_j \tau_j}$,

$$\frac{\partial N_{j-1}^\star(N)}{\partial N} \geqslant \frac{\partial N_j^\star(N)}{\partial N} \geqslant 0$$

□

The following lemma solves Problem (B.6) where there are only HVs.

Lemma 24. The optimal solution to Problem (B.6) can be expressed as:

$$\lambda_{H,j}^\ddagger = \left[\mu_j - \sqrt{\frac{\mu_j}{\tau_j}} \cdot \frac{-d + \sqrt{\Delta}}{2(1 - \gamma P_H/r)(\sum_{j=1}^J \sqrt{\mu_j \tau_j})}\right] \mathbf{1}_{j \leqslant J} \tag{B.30}$$

where $d = \sum_{j=1}^J (1 - (1 - \gamma P_H/r)\mu_j \tau_j)$, $\Delta = d^2 + 4(1 - \gamma P_H/r)(\sum_{j=1}^J \sqrt{\mu_j \tau_j})^2$, and J is also derived from Problem (B.6).

Proof. Proof of Lemma 24. Recall that Problem (B.6) is

$$\begin{aligned} \max_{\lambda_{H,j} \in [0,\mu_j), N_H \geqslant 0} \quad & \sum_{j=1}^L \tau_j \lambda_{H,j} \\ \text{s.t.} \quad & \sum_{j=1}^L (\tau_j + \frac{1}{\mu_j - \lambda_{H,j}})\lambda_{H,j} = N_H \\ & \gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j} = r N_H \end{aligned} \tag{24}$$

which is equivalent to

$$\begin{aligned} \max_{\lambda_{H,j} \in [0,\mu_j)} \quad & \sum_{j=1}^L \tau_j \lambda_{H,j} \\ \text{s.t.} \quad & \sum_{j=1}^L (\tau_j + \frac{1}{\mu_j - \lambda_{H,j}})\lambda_{H,j} = \frac{\gamma P_H \sum_{j=1}^L \tau_j \lambda_{H,j}}{r} \end{aligned} \tag{B.31}$$

Notice that we must have $\gamma P_H > r$; otherwise, there are no feasible solutions.

We can use the method of Lagrange multipliers to solve Problem (B.31) and construct the Lagrangian function:

$$\mathcal{L}(\{\lambda_{H,j}\}_{j=1}^{L}, \theta) = -\sum_{j=1}^{L} \tau_j \lambda_{H,j} + \theta \left(\sum_{j=1}^{L} (\tau_j + \frac{1}{\mu_j - \lambda_{H,j}}) \lambda_{H,j} - \frac{\gamma P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j}}{r} \right) - \sum_{j=1}^{L} \phi_j \lambda_{H,j}$$

where θ and ϕ_j are Lagrange multipliers. The Kuhn-Tucker conditions are:

- Stationarity:

$$-\tau_j + \theta(\tau_j + \frac{\mu_j}{(\mu_j - \lambda_{H,j})^2} - \frac{\gamma P_H \tau_j}{r}) - \phi_j = 0$$

- Primal feasibility:

$$\sum_{j=1}^{L} \left(\tau_j + \frac{1}{\mu_j - \lambda_{H,j}} \right) \cdot \lambda_{H,j} = \frac{\gamma P_H \sum_{j=1}^{L} \tau_j \lambda_{H,j}}{r}$$

$$\lambda_{H,j} \geqslant 0, j \in \{1, \ldots, L\}$$

- Dual feasibility:

$$\phi_j \geqslant 0, j \in \{1, \ldots, L\}$$

- Complementary slackness:

$$\phi_j \lambda_{H,j} = 0, j \in \{1, \ldots, L\}$$

We use a superscript $\ddagger$ to denote an optimal solution to Problem (B.31). Because of the complementary slackness, we consider two cases: either $\lambda_{H,j}^{\ddagger} > 0$ or $\lambda_{H,j}^{\ddagger} = 0$. First, if $\lambda_{H,j}^{\ddagger}$ is positive at some location j, then $\phi_j^{\ddagger} = 0$, and the optimal $\lambda_{H,j}^{\ddagger}$ and $\theta^{\ddagger}$ must satisfy:

$$\tau_j - \theta^{\ddagger}(\tau_j + \frac{\mu_j}{(\mu_j - \lambda_{H,j}^{\ddagger})^2} - \frac{\gamma P_H \tau_j}{r}) = 0$$

And we have:

$$(\mu_j - \lambda_{H,j}^{\ddagger})^2 = \frac{\theta^{\ddagger} \mu_j}{(1 - \theta^{\ddagger} + \theta^{\ddagger} \gamma P_H / r) \tau_j}$$

$$\lambda_{H,j}^{\ddagger} = \mu_j - \sqrt{\frac{\theta^{\ddagger} \mu_j}{(1 - \theta^{\ddagger} + \theta^{\ddagger} \gamma P_H / r) \tau_j}} \quad \text{because } \lambda_{H,j}^{\ddagger} < \mu_j$$

Now by considering the equality constraint:

$$\sum_{j=1,\lambda_{H,j}^{\ddagger}>0}^{L}(\tau_j+\frac{1}{\mu_j-\lambda_{H,j}^{\ddagger}})\lambda_{H,j}^{\ddagger}=\frac{\gamma P_H\sum_{j=1}^{L}\tau_j\lambda_j^{\ddagger}}{r}$$

$$\Longrightarrow\sqrt{\frac{\theta^{\ddagger}}{(1-\theta^{\ddagger}+\theta^{\ddagger}\gamma p/r)}}=\frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1,\lambda_{H,j}^{\ddagger}>0}^{L}\sqrt{\mu_j\tau_j})} \tag{B.32}$$

where $d=\sum_{j=1,\lambda_{H,j}^{\ddagger}>0}^{L}(1-(1-\gamma P_H/r)\mu_j\tau_j)$ and $\Delta=d^2+4(1-\gamma P_H/r)(\sum_{j=1,\lambda_{H,j}^{\ddagger}>0}^{L}\sqrt{\mu_j\tau_j})^2$. Notice that $\sqrt{\frac{\theta^{\ddagger}}{(1-\theta^{\ddagger}+\theta^{\ddagger}\gamma P_H/r)}}\neq\frac{-d-\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1,\lambda_{H,j}^{\ddagger}>0}^{L}\sqrt{\mu_j\tau_j})}$, because it is optimal to choose the larger $\lambda_{H,j}^{\ddagger}$. Therefore, when $\lambda_{H,j}^{\ddagger}>0$, we must have:

$$\lambda_{H,j}^{\ddagger}=\mu_j-\sqrt{\frac{\mu_j}{\tau_j}}\cdot\frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1,\lambda_{H,j}^{\ddagger}>0}^{L}\sqrt{\mu_j\tau_j})}$$

Second, if $\lambda_{H,j}^{\ddagger}$ is zero at some location j, by the Kuhn-Tucker conditions, we have $\phi^{\ddagger}\geqslant 0$ and:

$$\tau_j-\theta^{\ddagger}(\tau_j+\frac{1}{\mu_j}-\frac{\gamma P_H\tau_j}{r})\leqslant 0$$

$$\Longleftrightarrow\mu_j\tau_j\leqslant\frac{\theta^{\ddagger}}{1-\theta^{\ddagger}+\theta^{\ddagger}\gamma P_H/r}$$

meaning that $\lambda_{H,j}^{\ddagger}>0$ if and only if $\mu_j\tau_j>\frac{\theta^{\ddagger}}{1-\theta^{\ddagger}+\theta^{\ddagger}\gamma P_H/r}$. Recall that we assumed that $\mu_j\tau_j\geqslant\mu_{j+1}\tau_{j+1}$ for any $j\in\{1,\ldots,L-1\}$, so there exists $J\in\{1,\ldots,L\}$ such that $\lambda_{H,j}^{\ddagger}>0$ if and only if $j\geqslant J$. We then are able to express the optimal solution to Problem (B.6) as:

$$\lambda_{H,j}^{\ddagger}=\left[\mu_j-\sqrt{\frac{\mu_j}{\tau_j}}\cdot\frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})}\right]\mathbf{1}_{j\leqslant J}$$

where $d=\sum_{j=1}^{J}(1-(1-\gamma P_H/r)\mu_j\tau_j)$, $\Delta=d^2+4(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})^2$. And J can be found by solving:

$$J=\max\{j|\mu_j\tau_j>\left[\frac{-d+\sqrt{\Delta}}{2(1-\gamma P_H/r)(\sum_{j=1}^{J}\sqrt{\mu_j\tau_j})}\right]^2,\ j\in\{1,\ldots,L\}\} \tag{B.33}$$

□

The following lemma proves the sufficient conditions such that the lower bound (B.3c) or the upper bound (B.3d) of Problem ($\mathcal{M}'$) is binding.

Lemma 25. In Problem ($\mathcal{M}'$), Constraint (B.3c) is binding if N_A is sufficiently large (i.e., $N_A \to \infty$) or if N_A is sufficiently small and γ is sufficiently large (i.e., $N_A \to 0$ and $\gamma \to 1^-$); Constraint (B.3d) is binding when γ and N_A are sufficiently small (i.e., $N_A \to 0$ and $\gamma \to 0$).

Proof. Proof of Lemma 25. When neither of constraints (B.3c) and (B.3d) is binding. By Equation (B.9), we know the interior solution is:

$$\lambda_j^{int} = [\mu_j - \sqrt{\frac{\mu_j}{\tau_j}}\sqrt{\frac{\hat{r}}{p-\hat{r}}}]\mathbf{1}_{\mu_j\tau_j \geqslant \hat{r}/(p-\hat{r})} \quad \forall j \in \{1, \ldots, L\}$$

We can see that the interior solution is irrelevant with N_A.

Constraint (B.3c) **is binding** By Lemma 20 and Lemma 17, the optimal objective value of Problem (B.5), $\sum_{j=1}^{L} \tau_j \lambda_j^{\dagger}$, strictly increases in $\gamma \in (\tilde{\gamma}, \infty)$, strictly decreases in $N_A \in [0, \widetilde{N}_A)$ and strictly increases in $N_A \in [\widetilde{N}_A, \infty)$, where $\tilde{\gamma}$ and $\widetilde{N}_A$ are defined in Lemma 20. It is clear that $\sum_{j=1}^{L} \tau_j \lambda_j^{\dagger} \to \sum_{j=1}^{L} \tau_j \mu_j$ as $N_A \to \infty$. Thus, we must have $\sum_{j=1}^{L} \tau_j \lambda_j^{int} \leqslant \sum_{j=1}^{L} \tau_j \lambda_j^{\dagger}$ when N_A is sufficiently large.

In addition, suppose $\gamma = 1$ and $N_A = 0$, then $\sum_{j=1}^{L} \tau_j \lambda_j^{\dagger}$ can be derived from Problem (B.6). By Lemma 24,

$$\lambda_j^{\dagger} = \left[\mu_j - \sqrt{\frac{\mu_j}{\tau_j}} \cdot \frac{-d + \sqrt{\Delta}}{2(1 - P_H/r)(\sum_{j=1}^{J} \sqrt{\mu_j \tau_j})}\right] \mathbf{1}_{j \leqslant J}$$

where $d = \sum_{j=1}^{J}(1 - (1 - P_H/r)\mu_j\tau_j)$, $\Delta = d^2 + 4(1 - P_H/r)(\sum_{j=1}^{J} \sqrt{\mu_j\tau_j})^2$.

Therefore, $\sum_{j=1}^{L} \tau_j \lambda_j^{int} < \sum_{j=1}^{L} \tau_j \lambda_j^{\dagger}$ if and only if

$$\frac{-d + \sqrt{\Delta}}{2(1 - P_H/r)(\sum_{j=1, \lambda_j^{\dagger} > 0}^{L} \sqrt{\mu_j \tau_j})} < \sqrt{\frac{\hat{r}}{p - \hat{r}}} = \sqrt{\frac{r(P_A - c_A)}{(P_H - r)(P_A - c_A)}} = \sqrt{\frac{r}{P_H - r}}$$

Let us prove the above inequality backward:

$$
\begin{aligned}
&\frac{-d+\sqrt{\Delta}}{2(1-P_H/r)(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j})} < \sqrt{\frac{r}{P_H-r}} \\
\Longleftrightarrow & d-\sqrt{\Delta} < 2\sqrt{\frac{P_H-r}{r}}(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j}) \\
\Longleftrightarrow & d^2-2\sqrt{\Delta}d+\Delta < \frac{4(P_H-r)}{r}(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j})^2 \\
\Longleftrightarrow & 2d^2+4(1-P_H/r)(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j})^2-2\sqrt{\Delta}d < \frac{4(P_H-r)}{r}(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j})^2 \\
\Longleftrightarrow & d^2+2(1-P_H/r)(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j})^2-\sqrt{\Delta}d < \frac{2(P_H-r)}{r}(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j})^2 \\
\Longleftrightarrow & d^2-\sqrt{\Delta}d < \frac{4(P_H-r)}{r}(\sum_{j=1,\lambda_j^\dagger>0}^{L}\sqrt{\mu_j\tau_j})^2 \\
\Longleftrightarrow & d^2-\sqrt{\Delta}d < d^2-\Delta \\
\Longleftrightarrow & d > \sqrt{\Delta}
\end{aligned}
$$

where the last inequality is true because $P_H > r$ and the definitions of d and Δ in Equation (B.32).

Thus, if $\gamma = 1$ and $N_A = 0$, Constraint (B.3c) is binding. And because λ_j^{int} and $\lambda_j^\dagger$ are continuous in γ, these imply that Constraint (B.3c) is binding if N_A is sufficiently small and γ is sufficiently large (i.e., $N_A \to 0$ and $\gamma \to 1^-$)

Hence, in Problem ($\mathcal{M}'$), Constraint (B.3c) is binding if N_A is sufficiently large (i.e., $N_A \to \infty$) or if N_A is sufficiently small and γ is sufficiently large (i.e., $N_A \to 0$ and $\gamma \to 1^-$).

Constraint (B.3d) **is binding** Problem (B.6) can be considered as Problem (B.5) with $N_A = 0$, so by Lemma 20, $N_H^\ddagger$ strictly increases in $\gamma \in (\tilde{\gamma}, \infty)$, and for any $\gamma \in [0, \tilde{\gamma}]$, $N_H^\ddagger = 0$. Also, since $h(N)$ is strictly increasing in N, by Lemma 17, $\sum_{j=1}^{L}\tau_j\lambda_j^\ddagger$ strictly increases in

$\gamma \in (\tilde{\gamma}, \infty)$ and in N_A. And clearly, when $\gamma = 0$ and $N_A = 0$, $\sum_{j=1}^{L} \tau_j \lambda_j^{\ddagger} = 0$. And because λ_j^{int} and $\lambda_j^{\ddagger}$ are continuous, Constraint (B.3d) is binding if γ and N_A are sufficiently small and close to 0.

□

B.3 Simulation Study

This section describes how we process the dataset and design the simulation in detail. Please refer to Appendix B.4 for other auxiliary simulation results such as the omitted detailed maps in the robustness check in Section 3.6.3.

B.3.1 Data processing

To estimate the parameters in our model, we use the New York City Open Data platform to access the historical record of High-Volume For-Hire Vehicle (HVFHV) data.[1] For each trip in NYC, this data contains the origin, destination, and request time stamp for Lyft, Uber, and Via (the three leading ride-hailing platforms in NYC in 2020). For a more balanced demand distribution across the city (inflow and outflow are similar within zones), we consider the trip data between 11 AM and 1 PM during workdays in January 2020 (the month before the coronavirus led to a significant drop in demand), which corresponds to a total of 1,093,431 recorded trips.

Locations, zones and travel time For privacy reasons, the exact origin and destination of each trip are unavailable, but we have access to "taxi zones". NYC is divided into 257 such zones, which are chosen based on historical and demographic criteria. Examples of zones include "Times Square" and "JFK Airport", and the shape of the zones can be visualized in

[1] `https://data.cityofnewyork.us/Transportation/2020-High-Volume-FHV-Trip-Records/yrt9-58g8`, last accessed: 2024-05-29.

Figure 3.6. To generate exact locations and travel time, we use OpenStreetMap,[2] to obtain the road network of NYC, which can be visualized in Figure 3.6c and assume that each trip starts and end in a uniformly random intersection from the origin and destination zone, respectively. Specifically, in the origin zone of a request, we uniformly select one of the nodes in the zone at random as the origin node of this request. Similarly, in the destination zone of a request, we uniformly select one of the nodes in the zone at random as the destination node of this request. In addition, the travel time between each pair of nodes is available in the geospatial dataset, so we use the travel time between the origin node and the destination node as the travel time of a request. Notice that because the travel time in the geospatial dataset is computed by using the maximum speed, the travel time between each pair of nodes is much shorter than that in practice. To remedy this bias, we scale up the travel time between each pair of nodes by 3.

Demand density To obtain the demand density, we first calculated the hourly trips per zone in our trip dataset (i.e., the NYC HVFHV dataset). The demand density in a zone is then computed by dividing the hourly trips in that zone by its area,[3] see Figure 3.6a. We consider a zone with a higher demand density as a high-demand zone and a zone with a lower demand density as a low-demand zone. Let μ_k denote the demand density in zone k.

Pay ratio and reserved earning. For the pay ratio, in line with practice, we assume $\gamma = 75\%$.

To set the HVs' reserved earning, r, we consider NYC's average High-Volume FHV utilization rate and the average earning of fully-utilized HVs. The TLC reports that NYC's average citywide utilization rate is approximately 60%.[4] And the average earning of fully-

[2] `https://osmnx.readthedocs.io/en/stable/`, last accessed: 2024-05-29.

[3] For the airports, we used an estimate of the "driving area" of 1km^2 rather than counting the entire airport area, because the driving area in an airport is usually much smaller than its actual area

[4] `https://www1.nyc.gov/assets/tlc/downloads/pdf/fhv_congestion_study_report.pdf`

utilized HVs can be determined using vehicle pay regulations imposed by the Taxi and Limousine Commission (TLC) of NYC.[5] In NYC, the TLC requires that the combined vehicle pay rates correspond to \$1.161 per mile and \$0.529 per minute as of March 2022. Assuming an average vehicle speed of 20 miles per hour, we obtain that the average earning of fully-utilized HVs is $\$1.161 \times 20 + \$0.529 \times 60 = \$54.96$. We, therefore, set $r = \$33$ per hour, so that $r = 60\% \times \$54.96$.

B.3.2 Simulation Description

We repeated the simulation five times and took the average of all the results. At each repetition, the simulation is run as the following.

B.3.2.1 Data Sampling

At each iteration, we randomly sample one million requests from the trip data with replacement. This helps us to repeat the experiment, avoid cyclic demand imbalance, and take the average without affecting the demand distribution. We assume that requests arrive sequentially, and the inter-arrival time between two consecutive requests is generated from an exponential distribution with a mean of $1/23,770$, which is equal to $(2 \times 23)/1,093,431$ since we only consider 2 hours (i.e., 11 AM — 1 PM) during each workday and there are 23 workdays in January 2020. Consequently, we have a sample $\{t_i, o_i, d_i, \tau_{o_i,d_i}\}_{i=1}^{10^6}$, where t_i denotes the request time of request i, o_i denotes the origin node of request i, d_i denotes the destination node of request i, and τ_{o_i,d_i} denotes the travel time from o_i to d_i. Note that $t_i < t_{i+1},\ \forall i \in \{1, \ldots, 10^6 - 1\}$. And let $z(o_i)$ and $z(d_i)$ denote the origin zone and the destination zone of request i.

[5] `https://www1.nyc.gov/site/tlc/about/vehicle-pay.page` lists the regulated vehicle pay rates.

B.3.2.2 Core Steps

The core steps of the simulation describe how the platform, vehicles, and customers behave during the simulation, and how to output the desired metrics, such as service levels, revenues, profits, and ETA. In the core steps, the number of vehicles, N_A and N_H, and the price rates are inputs. Let $\{l_j^i\}_{j=1}^{N_A+N_H}$ denote the location of vehicle j when request i arrives. Initially, the vehicles' locations are uniformly chosen from the nodes in the geospatial street network.

ETA We assume the platform processes the requests sequentially. When request i arrives at time t_i, the platform computes the ETA between each available vehicle and the origin node of the request (i.e., the ETA is equal to $\tau_{l_j^i,o_i}$ for an available vehicle j).

Customer utility We assume that a customer facing a decision to travel has three options: selecting the nearest available HV, selecting the nearest available AV, or canceling the trip altogether. This choice is determined by a utility model that takes into account the vehicle type, hourly pricing, travel time from the origin to the destination, and the estimated time of arrival (ETA). Specifically, for a customer i requesting a trip, the customer's utility $U_{i,j}$ when assigned a vehicle j is given by:

$$U_{i,j} \triangleq [a_0 + \theta \mathbf{1}_{\text{j is an AV}} - (P_{A,i}\mathbf{1}_{\text{j is an AV}} + P_{H,i}\mathbf{1}_{\text{j is an HV}})]\tau_{o_i,d_i} + a_1\tau_{l_j^i,o_i}$$

where $a_0 > 0$ represents the base utility of choosing an HV and reaching the intended destination, θ captures the difference in utility if opting for an AV, $P_{A,i}$ and $P_{H,i}$ are the price rates (to be optimized) shown for customer i for AVs and HVs respectively, and $a_1 < 0$ is the customer's sensitivity to waiting time before pickup.

The utility $U_{i,j}$ measures a consumer's surplus from riding vehicle j in dollars. If $U_{i,j} < 0$ for all the available vehicles, the consumer will cancel the request and leave the market. If at least one available vehicle j yields a positive $U_{i,j}$, the customer will choose the option that maximizes her utility. That is, she will select an available vehicle J_i with the highest

(non-negative) utility:

$$J_i = \underset{\text{j is available and } U_{i,j} \geq 0}{\arg\max} U_{i,j}$$

Note that this is equivalent to choosing the nearest available AV or HV depending on the prices and the customer's preference for the type of vehicle. And if a tie occurs, she will randomly choose one of the vehicles with the same utility.

Pricing and Prioritization In contrast to the queuing model, in the simulation, customers can choose the type of vehicle they want to take, and the platform can only influence customers' choices through a pricing policy. Due to the complexity of the system (mixed fleet and large state space), we limit ourselves to a manageable set of pricing policies, that allow us to potentially prioritize AVs or HVs, and to distribute AVs and HVs differently in the city. For example, our set of policies allows the platform to potentially keep the AVs in the high-demand areas (downtown) by preventing the use of AVs for rides that lead to low-demand areas.

We let the price rate $P_{A,i}$ (or $P_{H,i}$) offered to customer i to be equal to a base price that is identical for all the trips, plus an adjustment depending on whether the destination of a trip is Manhattan (i.e., the area with the highest demand). That is, for any customer i, the price rates are

$$P_{type,i} = P_{type,base} + \delta_{type}(2 \times \mathbf{1}_{d_i \text{ is in Manhattan}} - 1) \quad type \in \{A, H\}$$

where $P_{type,base}$ is the base price, δ_{type} is the adjustment, and $\mathbf{1}_{\text{destination of i is in Manhattan}}$ is an indicator function.

This four-parameter class of pricing policies is rich enough to give the platform the flexibility to both prioritize and allocate (to a particular region) any specific type of vehicle. For example, to increase the utilization of AVs, the platform can offer discounts on AV rides compared to HVs by lowering $P_{A,base}$ and δ_A. Greater discounts lead to higher utilization of AVs. Moreover, the adjustment part allows the platform to influence the allocation of

vehicles by incentivizing customers to select different types of vehicles depending on their destinations. For example, when $\delta_A < \delta_H$, there is an extra discount on AV rides for the customers whose destination is in Manhattan, meaning that the platform is trying to allocate more AVs in Manhattan.

Serve the requests If the customer decides to choose vehicle J_i, vehicle J_i will depart to pick up the customer and transport her to the destination. Then the platform collects a profit $(P_{A,i} - c_A)\tau_{o_i,d_i}$ if J_i is an AV or $(1-\gamma)P_H\tau_{o_i,d_i}$ if J_i is an HV. During the pickup and the service time, vehicle J_i is marked as unavailable. That is, from t_i to $t_i + \tau_{l^i_{J_i},o_i} + \tau_{o_i,d_i}$, vehicle J_i is unavailable and cannot be matched with any other requests. After completing the service, vehicle J_i will become available and wait for the next requests at node d_i. Let s_i denote whether request i is successfully served (i.e. $s_i = 1$ if request i is not canceled, and $s_i = 0$ if request i is canceled or there are no vehicles available at time t_i).

Relocation Without relocation, most drivers would end up in the same area given enough simulation time, potentially leading to large driver imbalances in the city. To prevent this, we consider a simple relocation policy that relocates vehicles after a period of time. Each vehicle has an exponential clock (with a mean of 2 hours). Once the clock's time is up, and the vehicle is idle, the vehicle is (instantaneously) relocated to a zone that is sampled according to the demand distribution. That is, the probability of choosing the zone is its hourly trips divided by the total hourly trips in the city, and, therefore, vehicles are more likely to reposition to high-demand locations (Afèche et al., 2023). Then the vehicle will be relocated to one of the nodes within the zone which is uniformly selected at random. After relocation, a new clock is generated, and the vehicle will restart its work.

Metrics Computation The above steps are summarized in Algorithm 5. The inputs are the number of AVs, the number of HVs, the base price rates, and the adjustments. The outputs may include any metrics we want to comprehend, such as profit, average ETA,

service levels, and average earnings of HVs. In particular, the profit is the total revenue from serving requests minus the earnings of HVs and the operational cost of AVs.

$$\text{profit} = \sum_{i=1}^{10^6} s_i \tau_{o_i,d_i} [(P_{A,i} - c_A)\mathbf{1}_{J_i \text{ is an AV}} + (1-\gamma) P_{H,i} \mathbf{1}_{J_i \text{ is a HV}}]$$

where $\mathbf{1}_{J_i \text{ is an AV}}$ and $\mathbf{1}_{J_i \text{ is an HV}}$ are indicator functions. The other important metrics are listed below:

overall average ETA = $(\sum_{i=1}^{10^6} \tau_{l^i_{J_i}, o_i})/10^6$

average ETA in zone k = $(\sum_{i=1}^{10^6} \tau_{l^i_{J_i}, o_i} \mathbf{1}_{z(o_i)=k})/(\sum_{i=1}^{10^6} \mathbf{1}_{z(o_i)=k})$

overall service level = $(\sum_{i=1}^{10^6} s_i)/10^6$

service level in zone k = $(\sum_{i=1}^{10^6} s_i \mathbf{1}_{z(o_i)=k})/(\sum_{i=1}^{10^6} \mathbf{1}_{z(o_i)=k})$

average earning of HVs (per hour per vehicle) = $(\gamma \sum_{i=1}^{10^6} s_i \tau_{o_i,d_i} P_{H,i} \mathbf{1}_{J_i \text{ is a HV}})/(N_H t_{10^6})$

utilization rate of AVs (HVs) = $(\sum_{i=1}^{10^6} s_i \tau_{o_i,d_i} \mathbf{1}_{J_i \text{ is and AV (HV)}})/(t_{10^6})$

B.3.2.3 Find the equilibrium $\widetilde{N}_H$

Recall that human vehicles are assumed to be strategic and join the market by gauging their earning rate against a reserved earning. As a result, the average earning of HVs should be equal to this reserved earning at equilibrium:

$$r = \frac{\gamma \sum_{i=1}^{10^6} s_i \tau_{o_i,d_i} P_{H,i} \mathbf{1}_{J_i \text{ is a HV}}}{\widetilde{N}_H t_{10^6}} \tag{B.34}$$

where $\widetilde{N}_H$ is the number of HVs at equilibrium.

Given N_A and the price rates, we need to find $\widetilde{N}_H$ in the simulation. To this end, we select several candidates of $\widetilde{N}_H$, run Algorithm 5 for each candidate, and record the average earning of HVs. We then use a linear interpolation method to find $\widetilde{N}_H$ such that Equation (B.34) is satisfied. We summarize this step in Algorithm 6.

The candidates set of $\widetilde{N}_H$ in this study is $\{100, 3000, 6000, 9000, 12000, 15000\}$ whose range is big enough to make sure there exists a valid $\widetilde{N}_H$ satisfying the wage equilibrium.

Algorithm 5 The core steps of the simulation model

Input: $N_A, N_H, P_{A,base}, P_{H,base}, \delta_A, \delta_H$

Output: Desired metrics such as profit, ETA, service levels, etc.

Initialization: Randomly select vehicles' initial locations from the nodes, and generate each vehicle's working hours from an exponential distribution.

for $i = 1, 2, \ldots, 10^6$ **do**

  Compute the price rates: $P_{t,i} = P_{t,base} + \delta_t(2 \times \mathbf{1}_{d_i \text{ is in Manhattan}} - 1) \quad t \in \{A, H\}$.

  for $j = 1, 2, \ldots, N_A + N_H$ **do**

    if j is available **then**

      Compute the utility:

$$U_{i,j} = [a_0 + \theta \mathbf{1}_{\text{j is an AV}} - (P_{A,i}\mathbf{1}_{\text{j is an AV}} + P_{H,i}\mathbf{1}_{\text{j is an HV}})]\tau_{o_i,d_i} + a_1\tau_{l^i_j,o_i}$$

    end if

  end for

  if there are at least one vehicle j at time t_i that is available and $U_{i,j} \geq 0$ **then**

    The customer chooses vehicle J_i with the highest utility $J_i = \arg\max_{\text{j is available and } U_{i,j} \geq 0} U_{i,j}$.

    Vehicle J_i is dispatched and unavailable from time t_i to $t_i + \tau_{l^i_{J_i},o_i} + \tau_{o_i,d_i}$.

    The platform earns a profit and $s_i = 1$.

    The location of vehicle J_i at time $t_i + \tau_{l^i_{J_i},o_i} + \tau_{o_i,d_i}$ becomes d_i.

  else the customer cancels the trip and $s_i = 0$.

  end if

  for $j = 1, 2, \ldots, N_A + N_H$ **do**

    if j is available and stays for more than its exponential clock **then**

      Vehicle j is relocated. Its starting point is randomly selected according to the hourly trips of each zone, and its new clock is generated from an exponential distribution.

    end if

  end for

end for

Algorithm 6 Find the equilibrium

1: **Input**: $N_A, P_{A,base}, P_{H,base}, \delta_A, \delta_H$, and a candidate set of $\widetilde{N}_H$

2: **Output**: $\widetilde{N}_H$

3: **for** each N_H in the candidate set **do**

4: run Algorithm 5

5: record the average earning of HVs

6: **end for**

7: Construct a function that maps N_H to the average earning of HVs by using linear interpolation.

8: Find $\widetilde{N}_H$ by a root-finding algorithm (e.g. The bisection method) such that:

$$r = \frac{\gamma \sum_{i=1}^{10^6} s_i \tau_{o_i,d_i} P_{H,i} \mathbf{1}_{J_i \text{ is a HV}}}{\widetilde{N}_H t_{10^6}}$$

B.3.2.4 Find the best price parameters

The final step is to find the best price parameters (i.e., the best base price rates and the best adjustments). As mentioned above, four parameters need to be optimized - the base price rates $P_{A,base}$ and $P_{H,base}$, and the adjustments δ_A and δ_H. To find the best parameters, we compare the profits with different parameter candidates. This step is summarized in Algorithm 7.

In the baseline setting described in Section 3.6.1.1 and Section 3.6.1.2, the candidate set of $P_{A,base}$ and $P_{H,base}$ is $\{72, 73, 74, 75, 76, 77, 78\}$, and the candidate set of δ_A and δ_H is $\{-1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0\}$. Note that the candidate sets are adjusted in the different settings in Section 3.6.3 to make sure that we can observe the profit will be lower if the parameters are too high or too low.

For different choices of N_A, we repeated the entire experiment 5 times and took the average of all the results. The values of N_A we test in this study are {0, 2000, 4000, 6000,

Algorithm 7 Find the best parameters

1: **Input**: N_A and the candidate sets of price parameters.
2: **Output**: the best base price rates $P_{A,base}$ and $P_{H,base}$, and the best adjustments δ_A and δ_H.
3: **for** each $P_{A,base}, P_{H,base}, \delta_A, \delta_H$ in the candidate sets **do**
4: run Algorithm 6 to find the corresponding equilibrium $\widetilde{N}_H$.
5: run Algorithm 5 with $\widetilde{N}_H$ and record the profit at equilibrium.
6: **end for**
7: Find the best parameters with the highest profit at equilibrium.

8000, 10000, 12000}. Notice that in all the simulation results in this study, the zones with average hourly trips less than 0.2 are removed as outliers. The data points in these zones are so few that their noise is large. As a result, four zones are removed.

B.3.2.5 Alternative relocation method for HVs

In reality, HVs may be more strategic and have their own way to relocate themselves and maximize their earnings. To test our results in such a situation, we apply the method introduced by Braverman et al. (2019) and solve the linear programming (LP) problem in their lemma 1 to find the relocation destinations of HVs. Notice that this method requires much more computational time than our simple relocation method explained above since there are multiple linear programming problems to be solved during the simulation. So we only test it as a robustness check rather than apply it in the baseline model (see Section 3.6.3).

Similar to the original relocation method, each vehicle still has an exponential clock (with a mean of 2 hours). Once the clock's time is up, and the vehicle is idle, the vehicle is (instantaneously) relocated to a zone. However, in this alternative relocation method, the destination zone is not sampled according to the demand distribution. Specifically, instead of sampling a zone according to the demand distribution, we sample a zone according to

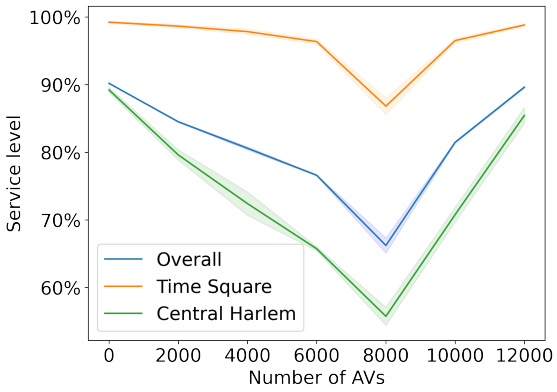
the solution of the LP problem (i.e., the empty-car routing policy Q defined in Braverman et al. (2019)). This policy Q is a matrix where an element Q_{ij} represents the transition probability for an HV in zone i to be relocated to zone j. In addition, to construct the LP problem, we need to find the demand rates for HVs and the average travel time of demand for HVs. This is challenging in our setting because the demand rates and the average travel time of demand for HVs are exogenously given in Braverman et al. (2019), but these are endogenously determined in our simulation. To this end, after every 2 hours (simulation time), we count the number of trips served by HVs from zone i to zone j and compute the average travel time of the trips served by HVs from zone i to zone j, for any i, j. We then use them as the demand rates and the travel time to solve the LP problem. The solution Q is used as the empty-HV relocation policy during the next two-hour time window. In other words, at the beginning of each two-hour period, we use the simulated data from the last period to solve the LP problem and update the relocation policy for HVs in the current period. Whenever an HV needs to be relocated in the current period, we sample a zone according to the relocation policy Q; and then the HV will be randomly relocated to one of the nodes within this zone.

B.4 Auxiliary Simulation Results

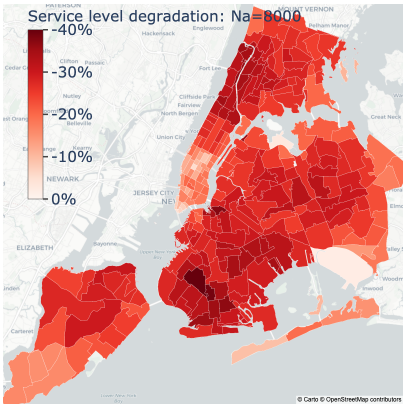
This section discusses some auxiliary results from our simulation study.

B.4.1 More results in robustness check (see Section 3.6.3)

Figures B.3 to B.14 present more detailed curves and maps about service levels in different settings, which are omitted in the robustness check in Section 3.6.3.

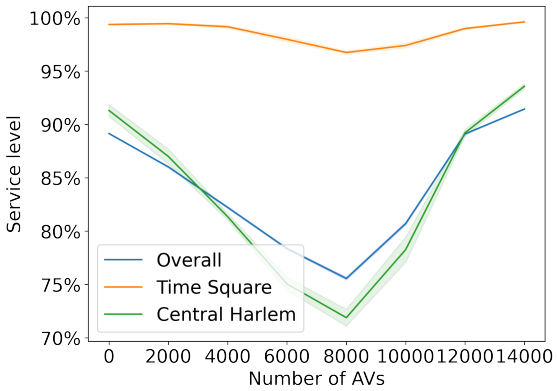


(a) Service level.

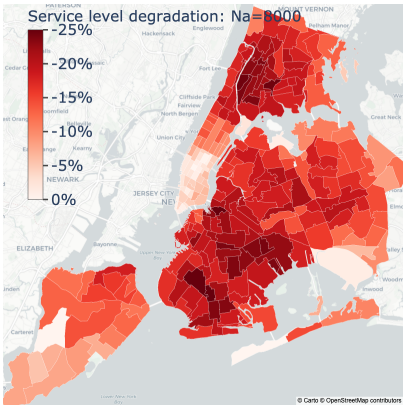


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.3: Robustness check: $\theta = 60$

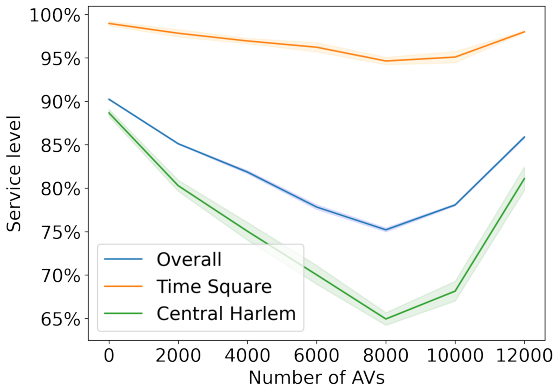


(a) Service level.

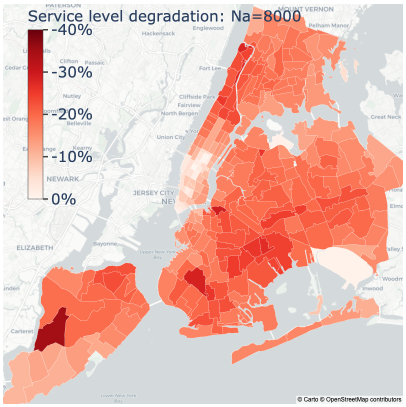


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.4: Robustness check: $\theta = 10$

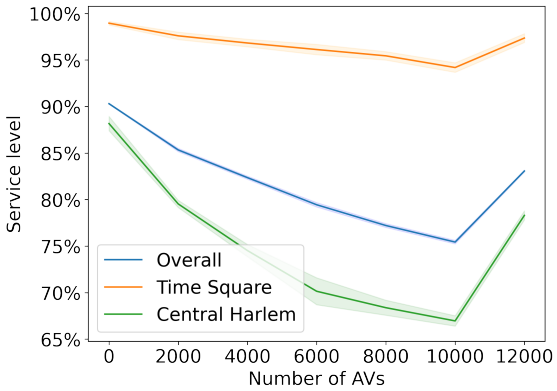


(a) Service level.

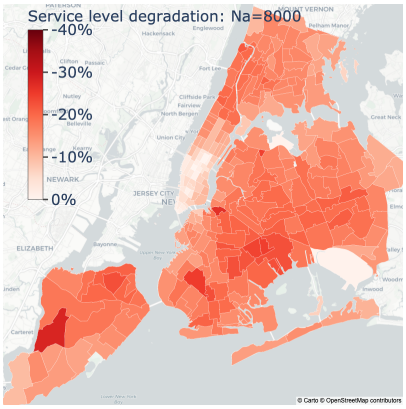


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.5: Robustness check: $\theta = -20$

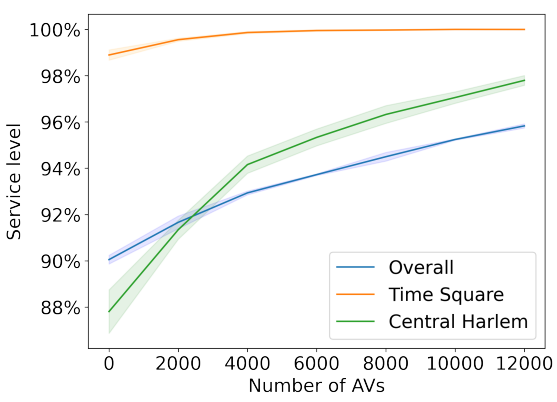


(a) Service level.

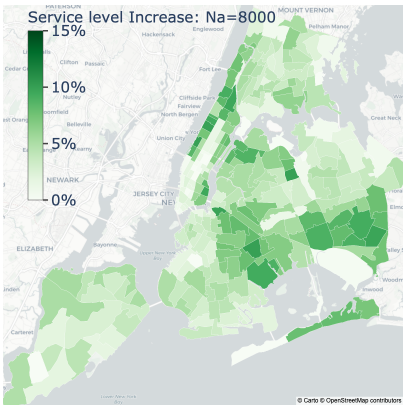


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.6: Robustness check: $\theta = -40$

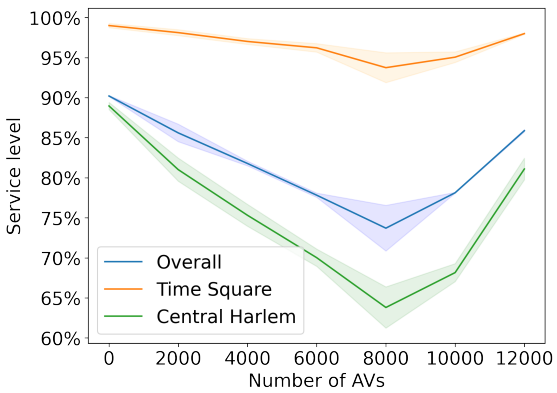


(a) Service level.

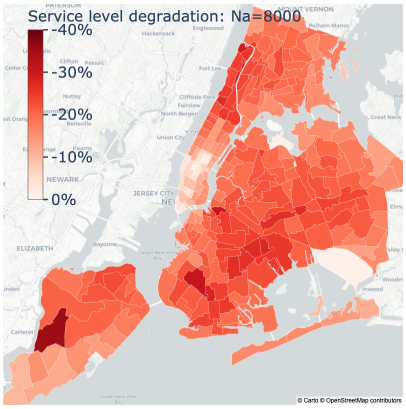


(b) Service level increases when 8,000 AVs are introduced.

Figure B.7: Robustness check: $\theta = -60$

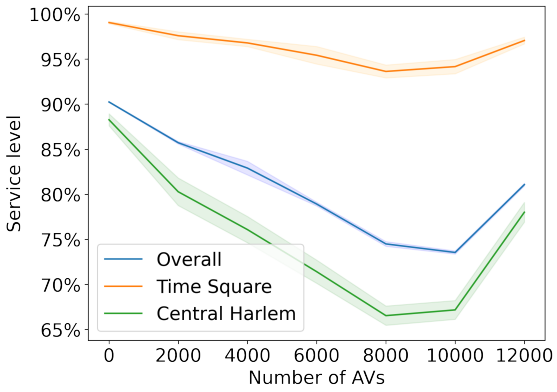


(a) Service level.

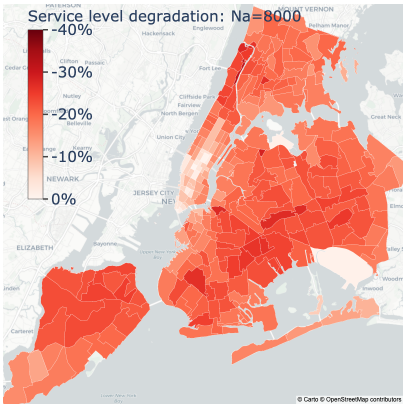


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.8: Robustness check: $c_A = 10$

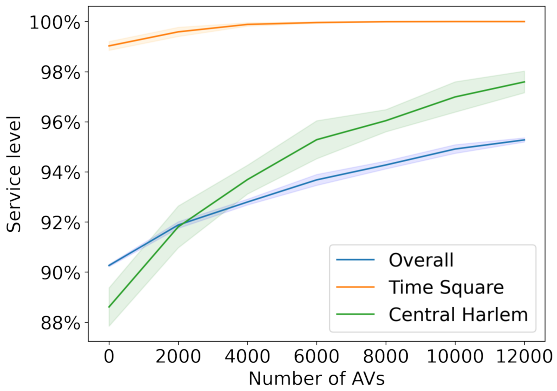


(a) Service level.

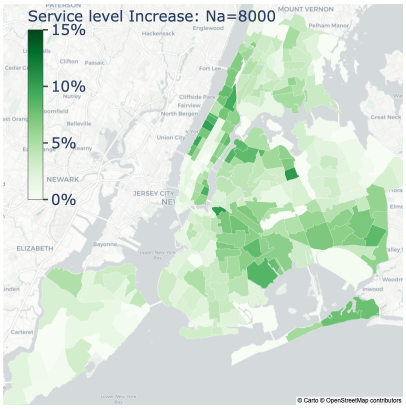


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.9: Robustness check: $c_A = 40$

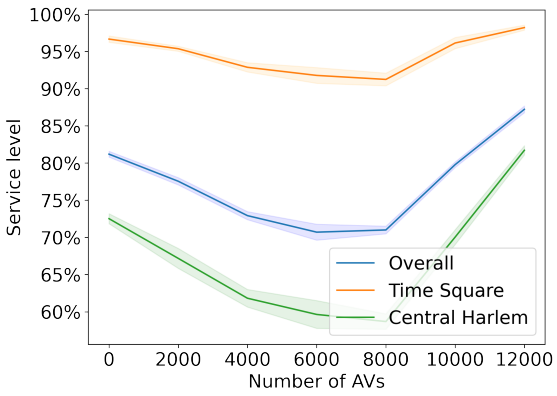


(a) Service level.

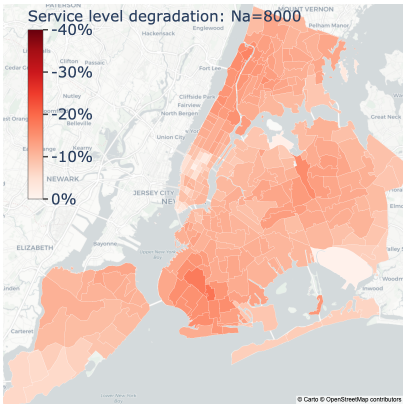


(b) Service level increases when 8,000 AVs are introduced.

Figure B.10: Robustness check: $c_A = 60$

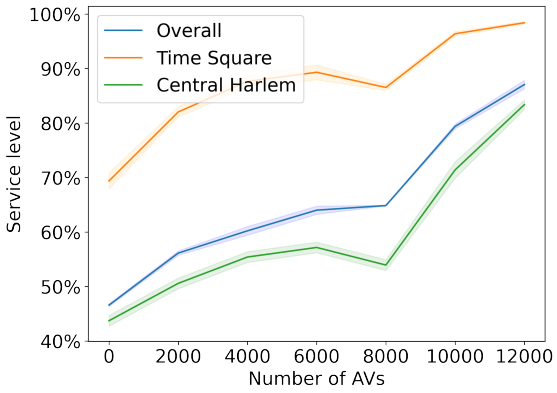


(a) Service level.

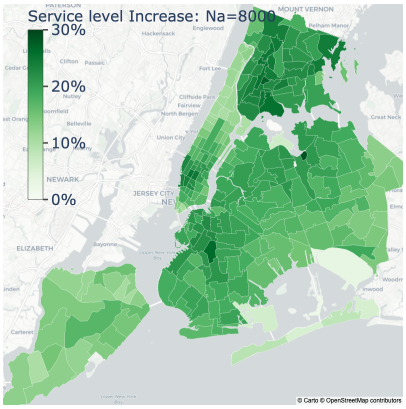


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.11: Robustness check: $\gamma = 60\%$

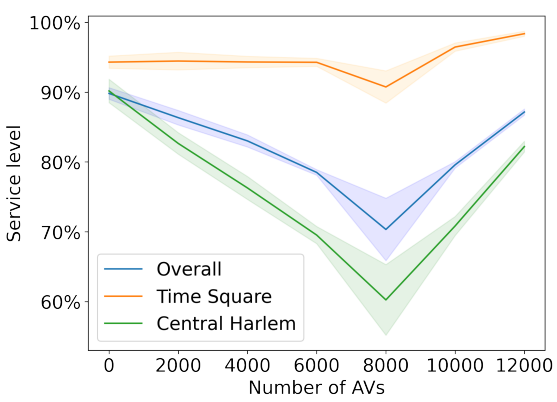


(a) Service level.

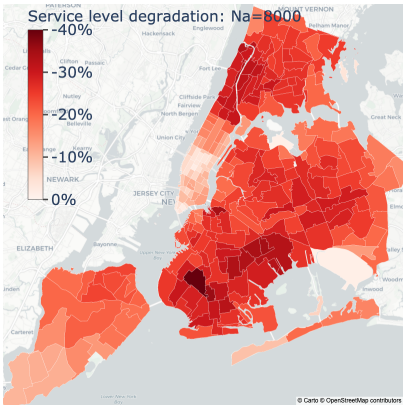


(b) Service level increases when 8,000 AVs are introduced.

Figure B.12: Robustness check: $\gamma = 50\%$

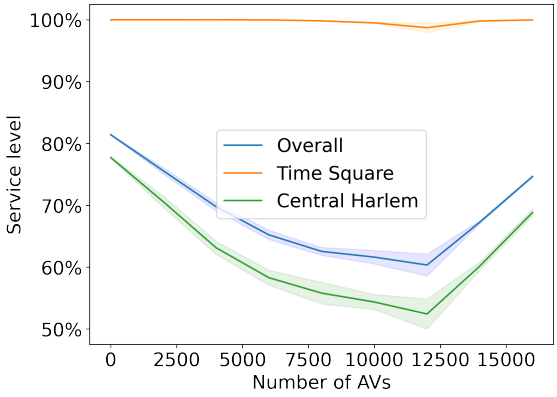


(a) Service level.

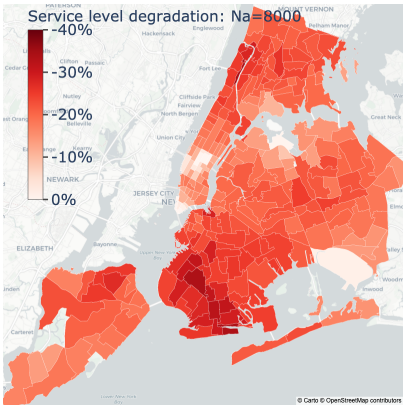


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.13: Robustness check: Relocation of HVs by Braverman et al. (2019).



(a) Service level.

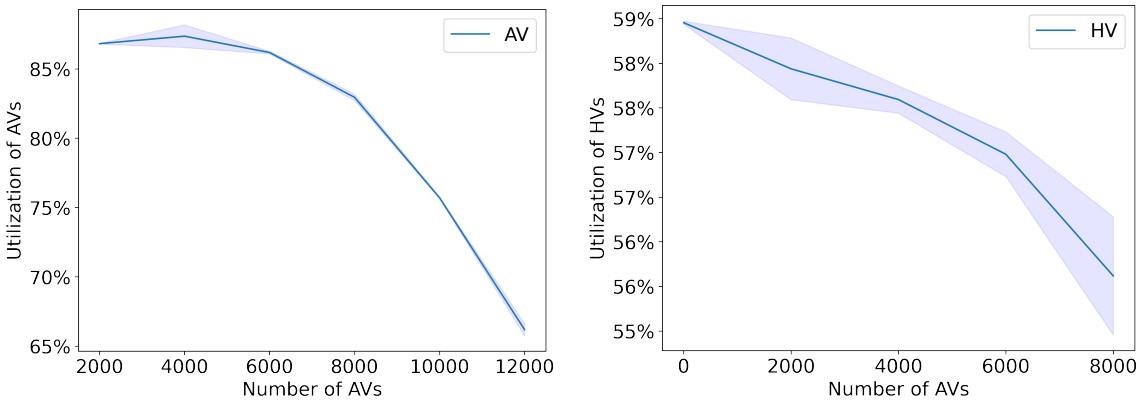


(b) Service level degradation if 8,000 AVs are introduced.

Figure B.14: Robustness check: the degradation of service levels when we use the dataset for the morning rush hour (7 AM - 9 AM) during the workdays in January 2020.

B.4.2 Optimality of Prioritizing AVs

To show the optimality of prioritizing AVs, we report the utilization of vehicles (cf., Figure B.15) and the best price parameters (cf., Figure B.16 and Figure B.17) obtained from our experiment with the baseline setting. As we can see the utilization of AV is much larger than the utilization of HVs, meaning that it is optimal to prioritize AVs.[6] The is consistent with the implication from price rates, as we can see that the best price rates for AVs are lower than the best price rates for HVs. In addition, Figure B.17 shows that the best price adjustment of AVs is also lower than the best price adjustment of HVs. This means the platform is encouraging customers to choose AVs when their destination is in Manhattan so that more AVs can be allocated in high-demand areas.

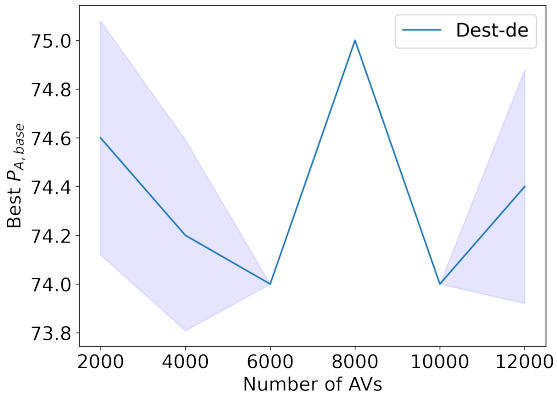


(a) The utilization rate of AVs.

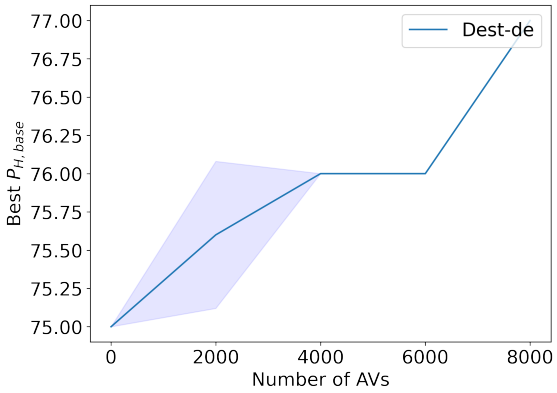
(b) The utilization rate of HVs.

Figure B.15: The utilization rate of vehicles.

[6]Notice that the utilization of HVs when $N_A = 0$ is close to 60% which aligns with our purpose when setting the reserved earning of HVs, r. See Appendix B.3.1.

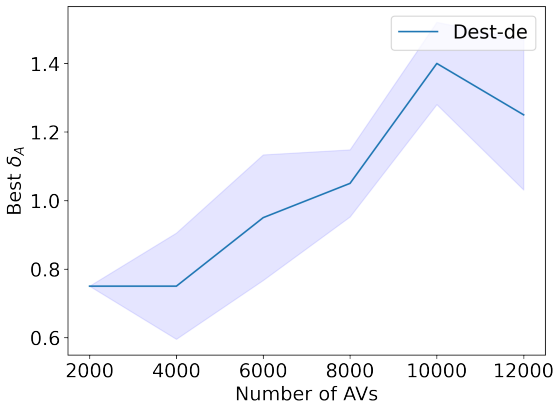


(a) The best base price rate of AVs $P_{A,base}$.

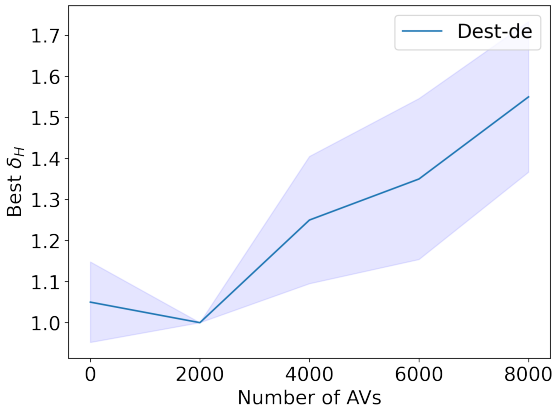


(b) The best base price rate of HVs $P_{H,base}$.

Figure B.16: The best base price rates in our experiment with the baseline setting.



(a) The best adjustment of AVs δ_A.



(b) The best adjustment of HVs δ_H.

Figure B.17: The best price adjustments in our experiment with the baseline setting.

APPENDIX C

Supply Prioritization in Hybrid Marketplaces

C.1 Proofs for Section 4.2

Proof. Proof of Lemma 3. We start from Problem (4.3):

$$\begin{aligned} \max_{R_P, R_F, N_F} \quad & R_P + (1-\gamma) R_F - C_P N_P \\ \text{s.t.} \quad & r N_F = \gamma R_F \\ & (R_P, R_F) \in \mathcal{AR}(N_P, N_F) \end{aligned}$$

Using the definition of $\mathcal{AR}(N_P, N_F)$, this is equivalent to:

$$\begin{aligned} \max_{R_P, R_F, N_F, \pi \in \Pi} \quad & R_P + (1-\gamma) R_F - C_P N_P \\ \text{s.t.} \quad & r N_F = \gamma R_F \\ & R_P = R_P^\pi(N_P, N_F) \\ & R_F = R_F^\pi(N_P, N_F) \end{aligned}$$

We can simplify this problem as R_F and R_F are fixed given N_F and π:

$$\begin{aligned} \max_{N_F, \pi \in \Pi} \quad & R_P^\pi(N_P, N_F) + (1-\gamma) R_F^\pi(N_P, N_F) - C_P N_P \\ \text{s.t.} \quad & r N_F = \gamma R_F^\pi(N_P, N_F) \end{aligned} \tag{C.1}$$

We can see that any solution π of Problem (2') must be feasible in the above Problem (C.1), since $N_F \in \mathcal{E}^\pi(N_P, \gamma) \implies r N_F = \gamma R_F^\pi(N_P, N_F)$. On the other hand, any solution

π, N_F of Problem (C.1) must be feasible in Problem (4.2), because $rN_F = \gamma R_F^{\pi}(N_P, N_F) \implies N_F \in \mathcal{E}^{\pi}(N_P, \gamma)$.

□

C.2 Proofs for Section 4.3

Proof. Proof of Lemma 4. Suppose there exists $\widetilde{R_F}$ feasible for Problem (4.4) that satisfies equal treatment and that leads to a positive number of flexible supply hours denoted by $\widetilde{N_F}$. And because of feasibility, we have that $r\widetilde{N_F} = \gamma\widetilde{R_F}$. Because of equal treatment, we have that there exists R_P such that $(R_P, \widetilde{R_F}) \in \mathcal{AR}(N_P, \widetilde{N_F})$ and $R_P\widetilde{N_F} = \widetilde{R_F}N_P$. The total revenue associated to this equal treatment solution can be no more than the maximal revenue of an equal treatment solution which by Assumption 1 and Assumption 2 is $\overline{R}(N_P + \widetilde{N_F})$. That is, we must have that $R_P + \widetilde{R_F} \leqslant \overline{R}(N_P + \widetilde{N_F})$. However, note that

$$R_P + \widetilde{R_F} = \widetilde{R_F}\frac{N_P}{\widetilde{N_F}} + \widetilde{R_F} = \frac{r}{\gamma}\widetilde{N_F}\frac{N_P}{\widetilde{N_F}} + \frac{r}{\gamma}\widetilde{N_F} = \frac{r}{\gamma}(\widetilde{N_F} + N_P).$$

Since $\widetilde{N_F} > 0$, we obtain the desired conclusion.

Conversely, suppose that there exists $N > N_P$ such that $rN \leqslant \gamma\overline{R}(N)$. Because there exists $M > 0$ such that $\forall N \geqslant 0,\ \overline{R}(N) \leqslant M$, we have $\lim_{N\to\infty} \overline{R}(N)/N = 0$. Therefore, with the continuity of $\overline{R}(\cdot)$, we can always find $\widehat{N} \geqslant N$ such that $r\widehat{N} = \gamma\overline{R}(\widehat{N})$ by the intermediate value theorem. Then, since $\frac{r}{\gamma}\widehat{N} = \overline{R}(\widehat{N}) \in \mathcal{AR}(\widehat{N})$, we can use Assumption 2 for N_P and $\widetilde{N_F} = \widehat{N} - N_P > 0$ to find $(R_P, \widetilde{R_F}) \in \mathcal{AR}(N_P, \widetilde{N_F})$ such that

$$R_P + \widetilde{R_F} = \frac{r}{\gamma}\widehat{N} \quad \text{and} \quad R_P\widetilde{N_F} = \widetilde{R_F}N_P.$$

Replacing R_P from the second expression into the first expression above, and using that $\widetilde{N_F} + N_P = \widehat{N}$ yields that $\widetilde{R_F} = \frac{r}{\gamma}\widetilde{N_F}$. In turn, $\widetilde{R_F}$ is feasible for Problem (4.10) and compatible with an equal treatment policy with positive flexible supply hours.

□

Proof. Proof of Proposition 12. In the following proof, we first show that when $\widetilde{N} > N_P$, Problem (4.4) is equivalent to Problem (C.2), then derive the result from Problem (C.2).

By Definition 2, Lemma 4, and $\widetilde{N} > N_P$, Problem (4.4) can be rewritten as:

$$\begin{aligned} \max_{N_F>0, R_P, R_F} \quad & R_P + (1-\gamma)R_F - C_P N_P \\ \text{s.t.} \quad & (R_P, R_F) \in \mathcal{AR}(N_P, N_F) \\ & \gamma R_F = r N_F \\ & N_P R_F = N_F R_P \end{aligned}$$

As $N_F > 0$, the two last constraints are equivalent to $R_P = N_P R_F / N_F$ and $R_F = r N_F / \gamma$. The problem is therefore equivalent to:

$$\begin{aligned} \max_{N_F>0} \quad & r N_P/\gamma + (1-\gamma) r N_F/\gamma - C_P N_P \\ \text{s.t.} \quad & r/\gamma(N_P, N_F) \in \mathcal{AR}(N_P, N_F) \end{aligned}$$

Reorganizing the objective, this is the same as:

$$\begin{aligned} \max_{N_F>0} \quad & r N_P + (1-\gamma) r(N_F + N_P)/\gamma - C_P N_P \\ \text{s.t.} \quad & r/\gamma(N_P, N_F) \in \mathcal{AR}(N_P, N_F) \end{aligned}$$

With the change of variable $N = N_P + N_F$ and using Assumption 1 and Assumption 2, this is equivalent to:

$$\begin{aligned} \max_{N>N_P} \quad & r N_P + (1-\gamma) r N/\gamma - C_P N_P \\ \text{s.t.} \quad & r N/\gamma \in \mathcal{AR}(N) \end{aligned}$$

Suppose $\widehat{N}$ is the optimal solution of the above problem. As $\gamma < 1$, the coefficient of $\widehat{N}$ in the objective is positive, therefore it is optimal to have $\widehat{N} = \max\{N \mid r N/\gamma \in \mathcal{AR}(N)\}$. This implies that $r\widehat{N}/\gamma \leqslant \overline{R}(\widehat{N})$. Suppose that $r\widehat{N}/\gamma < \overline{R}(\widehat{N})$. Then, $\overline{R}(\widehat{N}) - r\widehat{N}/\gamma > 0$ and $n \to \overline{R}(n) - rn/\gamma$ is continuous in n, since $\overline{R}(\cdot)$ is continuous. Also, $R(n) - rn/\gamma$ goes to $-\infty$ when $n \to +\infty$, as $\sup(\overline{R}(N))$ is bounded. Therefore, by the intermediate value theorem,

there exists $N' > \widehat{N}$ such that $\overline{R}(N') - rN'/\gamma = 0$, which contradicts the definition of $\widehat{N}$. We conclude that we must have $r\widehat{N}/\gamma = \overline{R}(\widehat{N})$ and therefore our problem is equivalent to:

$$\begin{aligned} \max_{N>N_P} \quad & (1-\gamma)\overline{R}(N) + (r - C_P)N_P \\ \text{s.t.} \quad & rN = \gamma \overline{R}(N) \end{aligned} \tag{C.2}$$

Note that the constraint $N > N_P$ is not necessary since we already assumed $\widetilde{N} > N_P$, and $\widetilde{N}$ is the optimal solution.

To conclude, when $N_P < \widetilde{N}$, the solution of Problem (C.2) is always $\widetilde{N}$ and independent of N_P. Therefore, $N_F^E = \widetilde{N} - N_P > 0$. By Definition 2 and the equilibrium, we must have $R_P^E = rN_P/\gamma$ and $R_F^E = rN_F^E/\gamma = \overline{R}(\widetilde{N}) - R_P^E$.

And when $N_P \geqslant \widetilde{N}$, we know $N_F^E = 0$ as a result of Lemma 4. This implies that the optimal solution has to maximize the revenue of private supply given N_P, so we have $R_P^E = \overline{R}(N_P)$. Notice that $R_P^E = \overline{R}(N_P)$, $N_F^E = 0$ must be an equal treatment solution (i.e. $\overline{R}(N_P)/N_P \leqslant r/\gamma$). Otherwise, if $\overline{R}(N_P)/N_P > r/\gamma$, since $\overline{R}(N)$ is continuous and $\lim_{N\to\infty} \overline{R}(N) < \infty$, there exists $N > N_P$ such that $\overline{R}(N)/N = r/\gamma$ by the intermediate value theorem, which contradicts the definition of $\widetilde{N}$.

□

Proof. Proof of Theorem 6. Let $(N_P^E, N_F^E, R_P^E, R_F^E)$ denote an optimal solution of Problem (4.5). Suppose this equal treatment solution has hybrid supply (i.e. $N_P^E > 0$ and $N_F^E > 0$). Notice that N_P^E has to be less than $\widetilde{N}$ by Lemma 4. We want to show that there exists a private-supply-only solution which has a higher or equal profit than this solution.

When $\gamma \in [0, 1)$, the result follows from Proposition 12. By Proposition 12, we have $N_P^E < \widetilde{N}$, $N_F^E = \widetilde{N} - N_P^E$, and its profit must be equal to $(1-\gamma)\overline{R}(\widetilde{N}) + (r - C_P)N_P^E$. Thus, depending on the difference between r and C_P, we have a single-type solution with a profit that is not less than $(1-\gamma)\overline{R}(\widetilde{N}) + (r - C_P)N_P^E$. Specifically, if $r \leqslant C_P$, we can reset $N_P^E = 0$, $N_F^E = \widetilde{N}$, and the profit will be increased to $(1-\gamma)\overline{R}(\widetilde{N})$; and if $r > C_P$, we can reset $N_P^E = \widetilde{N}$, $N_F^E = 0$, and the profit will be increased to $(1-\gamma)\overline{R}(\widetilde{N}) + (r - C_P)\widetilde{N}$.

When $\gamma \geqslant 1$, due to Definition 2, if $N_P^E > 0,\ N_F^E > 0$, we have $R_P^E = rN_P^E/\gamma$. Thus,

$$\begin{aligned} & R_P^E + (1-\gamma)R_F^E - C_P N_P^E \\ = & (\frac{r}{\gamma} - C_P)N_P^E + (1-\gamma)R_F^E \\ \leqslant & (\frac{r}{\gamma} - C_P)N_P^E \qquad \text{because } \gamma \geqslant 1 \end{aligned}$$

If $\frac{r}{\gamma} < C_P$, it means any profit of an equal treatment solution with hybrid supply is negative. In this case, the firm could simply shut down and reset $N_P^E = N_F^E = 0$ to have a higher profit.

Suppose $\frac{r}{\gamma} > C_P$, then because $N_P^E < \widetilde{N}$,

$$(\frac{r}{\gamma} - C_P)N_P^E \leqslant (\frac{r}{\gamma} - C_P)\widetilde{N}$$

This means that we only need to show $(N_P = \widetilde{N}, R_P = r\widetilde{N}/\gamma, R_F = N_F = 0)$ is a feasible equal treatment solution. Because $r\widetilde{N}/\gamma = \overline{R}(\widetilde{N})$, $(r\widetilde{N}/\gamma, 0) \in \mathcal{AR}(\widetilde{N}, 0)$ by Assumption 1, it satisfies the definition of equal treatment in Definition 2. Thus, $(N_P = \widetilde{N}, R_P = r\widetilde{N}/\gamma, R_F = N_F = 0)$ is an equal treatment solution with private supply only and leads to a higher profit than $(N_P^E, N_F^E, R_P^E, R_F^E)$.

Hence, there exists a single-supply solution whose profit is higher than or equal to any equal treatment solution with hybrid supply. □

Proof. Proof of Lemma 5.

Let $(R'_P, R'_F, \gamma', N'_F)$ be an optimal solution of Problem (4.6), there exists a feasible solution of Problem (4.7):

$$\widehat{R_P} = R'_P, \widehat{R_F} = R'_F, \widehat{N_F} = N'_F$$

such that Problem (4.6) and Problem (4.7) have the same objective value.

On the other hand, let $(\widehat{R_P}, \widehat{R_F}, \widehat{N_F})$ to be an optimal solution of Problem (4.7).

If $\widehat{R_F} > 0$ and $\widehat{N_F} > 0$, or $\widehat{R_F} = \widehat{N_F} = 0$, then there exists a feasible solution of Problem

(4.6):

$$R'_P = \widehat{R_P}, R'_F = \widehat{R_F}, N'_F = \widehat{N_F}, \gamma' = \begin{cases} \frac{r\widehat{N_F}}{\widehat{R_F}} & \widehat{N_F} > 0,\ \widehat{R_F} > 0 \\ 0 & \widehat{N_F} = 0 \end{cases}$$

such that Problem (4.7) and Problem (4.6) have the same objective value.

Suppose $\widehat{N_F} > 0$ but $\widehat{R_F} = 0$. we know that there exists $\widehat{R} \in \mathcal{AR}(N_P + N_F)$ such that $\widehat{R} = \widehat{R_P}$ due to Assumption 1. Also, by Assumption 2, there exists $(\widetilde{R_P}, \widetilde{R_F}) \in \mathcal{ET}(N_P, \widehat{N_F})$ such that $\widetilde{R_P} + \widetilde{R_F} = \widehat{R} = \widehat{R_P}$. This means $(\widetilde{R_P}, \widetilde{R_F}, \widehat{N_F})$ is also an optimal solution of Problem (4.7). By Definition 2, since $\widehat{N_F} > 0$, $(\widetilde{R_P}, \widetilde{R_F}) \in \mathcal{ET}(N_P, \widehat{N_F})$ implies either $N_P = 0$ or $\widetilde{R_P}/N_P = \widetilde{R_F}/\widehat{N_F}$, but only the latter case with $\widetilde{R_P} > 0$ is true. If $N_P = 0$ or $\widetilde{R_P} = 0$, then the optimal value of Problem (4.7) would be $-r\widehat{N_F} < 0$, which is impossible since $\widehat{N_F} = 0$ is always feasible and the objective value can be increased to 0. Therefore, we must have $N_P > 0$ and $\widetilde{R_P} > 0$. This implies $\widetilde{R_F} > 0$. Thus, $(\widetilde{R_P}, \widetilde{R_F}, \widehat{N_F})$ is also optimal with $\widetilde{R_F} > 0$ for Problem (4.7), then we can construct a feasible solution of Problem (4.6):

$$R'_P = \widehat{R_P}, R'_F = \widehat{R_F}, N'_F = \widehat{N_F}, \gamma' = \frac{r\widehat{N_F}}{\widetilde{R_F}}$$

such that Problem (4.7) and Problem (4.6) have the same objective value. We conclude that Problem (4.7) and Problem (4.6) are equivalent.

□

Proof. Proof of Theorem 7.

Let $(N_F^\star, R_P^\star, R_F^\star)$ be an optimal solution of Problem (4.7). Let $N'_F = N_F^\star$. By Assumption 1 and Assumption 2, we know that there exists R'_P, R'_F such that $(R'_P, R'_F) \in \mathcal{ET}(N_P, N'_F)$ and $R'_P + R'_F = R_P^\star + R_F^\star$. In addition, we can simply set

$$\gamma' = \begin{cases} \frac{rN'_F}{R'_F} & N'_F > 0 \\ 0 & N'_F = 0 \end{cases}$$

Note that if $N'_F > 0$ we must have $R'_F > 0$, otherwise, we would have $R'_P = 0$ by Definition 2. This means $R_P^\star + R_F^\star = R'_P + R'_F = 0$ and the objective value of $(N_F^\star, R^\star)$ in Problem (4.7) is

negative and lower than the objective value of $(N_F = 0, R_F = 0)$ which is 0. This contradicts the optimality of $(N_F^\star, R_P^\star, R_F^\star)$ for Problem (4.7).

Thus, $(R'_P, R'_F, N'_F, \gamma')$ is feasible in Problem (4.6), and its objective value is equal to the optimal value of Problem (4.6) by Lemma 5. Hence, $(R'_P, R'_F, N'_F, \gamma')$ is an optimal solution of Problem (4.6).

□

Proof. Proof of Corollary 2. We prove the cases $C_P < r$ and $C_P > r$ separately. In fact, the case with $C_P < r$ does not require γ to be variable, so we will relax this condition when proving the first part with $C_P < r$.

Suppose that $C_P < r$. For any γ, let (R'_P, R'_F, N'_P, N'_F) be an optimal solution of Problem (C.3) (by assuming γ is fixed):

$$\begin{aligned} \max_{R_P, R_F, N_P, N_F} \quad & R_P + (1-\gamma)R_F - C_P N_P \\ \text{s.t.} \quad & rN_F = \gamma R_F \\ & (R_P, R_F) \in \mathcal{AR}(N_P, N_F) \end{aligned} \tag{C.3}$$

We want to show $N'_F = 0$. For the sake of contradiction, suppose $N'_F > 0$ hence $\gamma R'_F > 0$. From this, we build another feasible solution $(R_P^\star, R_F^\star, N_P^\star, N_F^\star)$ that yields a larger objective. Let $N_P^\star = N'_P + N'_F$ and $N_F^\star = 0$, note that since $(R'_P, R'_F) \in \mathcal{AR}(N'_P, N'_F)$ by Assumption 1 we have that

$$R'_P + R'_F \in \mathcal{AR}(N'_P + N'_F) = \mathcal{AR}(N_P^\star),$$

and, therefore, using Assumption 1, we can find $(R_P^\star, R_F^\star) \in \mathcal{AR}(N_P^\star, 0)$ with $R_P^\star + R_F^\star = R'_P + R'_F$ and $R_F^\star = 0$. Note the solution $(R_P^\star, R_F^\star, N_P^\star, N_F^\star)$ also satisfies the equilibrium condition with $N_F^\star = 0$. For the objective we have the following:

$$\begin{aligned} R_P^\star + R_F^\star - C_P N_P^\star - \gamma R_F^\star &= R'_P + R'_F - C_P(N'_P + \gamma R'_F/r) \\ &> R'_P + R'_F - C_P N'_P - \gamma R'_F, \end{aligned}$$

where in the first equality we used that $N'_F = \gamma R'_F$ and in the inequality we used that $C_P < r$ and that $\gamma R'_F > 0$. The latter contradicts the optimality of (R'_P, R'_F, N'_P, N'_F). Notice that the above demonstration is valid for any possible value of γ including the case when γ is optimally chosen.

For the second case suppose that $C_P > r$. Additionally, assume that we can optimize over γ. Let $(R'_P, R'_F, N'_P, N'_F, \gamma')$ to be an optimal solution to Problem (4.8), and let N'_F be such that $rN'_F = \gamma' R'_F$. For the sake of contradiction, assume that $N'_P > 0$. Define $N^\star_P = 0$ and $N^\star_F = N'_P + N'_F$. Since $R'_P + R'_F \in \mathcal{AR}(N'_P + N'_F)$, then by Assumption 1, there exist $(R^\star_P, R^\star_F) \in \mathcal{AR}(0, N^\star_F)$ with $R^\star_F = R'_P + R'_F$ and $R^\star_P = 0$. Next we use $(R^\star_P, R^\star_F, N^\star_P, N^\star_F)$ to build a feasible solution. Let $\gamma^\star$ be defined by

$$\gamma^\star = \frac{rN^\star_F}{R^\star_F}. \tag{C.4}$$

Note that $R'_P + R'_F > 0$, otherwise, the objective value of $(R'_P, R'_F, N'_P, N'_F, \gamma')$ would be $-C_P N'_P < 0$, which is not optimal since we can always choose $N'_P = N'_F = 0$. Also, $N^\star_F$ satisfies the equilibrium condition because $\gamma^\star R^\star_F = r(N'_P + N'_F) = rN^\star_F$. Hence, $(R^\star_P, R^\star_F, N^\star_P, N^\star_F, \gamma^\star)$ is a feasible solution to Problem (4.8). This solution also achieves a higher objective value:

$$\begin{aligned} R^\star_P + (1-\gamma^\star)R^\star_F - C_P N^\star_P &= R'_P + R'_F - r(N'_P + N'_F) \\ &> R'_P + R'_F - C_P N'_P - rN'_F \\ &= R'_P + R'_F - C_P N'_P - \gamma' R'_F \\ &= R'_P + (1-\gamma')R'_F - C_P N'_P, \end{aligned}$$

where the strict inequality comes from $C_P > r$ and $N'_P > 0$. This contradicts the assumption that $(R'_P, R'_F, N'_P, N'_F, \gamma')$ is optimal. Hence, we should have $N'_P = 0$ in this case.

□

C.3 Proofs for Section 4.4.

Proof. Proof of Proposition 13. Let $(R_P^\star, R_F^\star, N_F^\star)$ be an optimal solution of Problem (4.3). Since $\gamma R_F^\star = rN_F^\star$ and $(R_P^\star, R_F^\star) \in \mathcal{AR}(N_P, N_F^\star)$, we have $(R_P^\star, R_F^\star) \in \mathcal{AR}(N_P, \gamma R_F^\star/r)$. Thus, $(\gamma R_F^\star/r, R_F^\star) \in \mathcal{D}$ and $R_F^\star$ is feasible in Problem (4.10). In addition, we must have $R_P^\star = \overline{R}_P(N_F^\star, R_F^\star)$, otherwise, we can always find another $R_P > R_P^\star$ such that $R_P, N_F^\star, R_F^\star$ is feasible in Problem (4.3) and produces an strictly higher objective value.

On the other hand, suppose R_F' is an optimal solution of Problem (4.10). Let $N_F' = \gamma R_F'/r$, $R_P' = \overline{R}_P(\gamma R_F'/r, R_F')$. Clearly, $rN_F' = \gamma R_F'$ and $(R_P', R_F') \in \mathcal{AR}(N_P, N_F')$, so (R_F', N_F', R_P') is feasible in Problem (4.3). Therefore, Problem (4.3) is equivalent to Problem (4.10).

□

The other results in this section require the following auxiliary lemma.

Lemma 26 (Properties of the optimal equal treatment solution)**.** Suppose Assumption 1 and Assumption 2 hold. Then, the private supply revenue of the optimal equal treatment solution is the maximum private supply revenue given the flexible supply hours and the flexible supply revenue:

$$R_P^E = \overline{R}_P(N_F^E, R_F^E).$$

This identity implies that the optimal equal treatment profit is $\text{Profit}(R_F^E)$.

Proof. Proof of Lemma 26.

1. If $N_P \geqslant \widetilde{N}$: Proposition 12 implies $R_P^E = \overline{R}(N_P) = \overline{R}_P(0,0)$.

2. If $N_P < \widetilde{N}$: consider the following sequence of equivalent definitions of $\overline{R}(N_P + N_F)$

$$\begin{aligned}
\overline{R}(N_P + N_F) &= \max \mathcal{AR}(N_P + N_F) \\
&= \max_{(R_P, R_F) \in \mathcal{AR}(N_P, N_F)} R_P + R_F \quad \text{by Assumption 1} \\
&= \max_{R_F \,|\, (N_F, R_F) \in \mathcal{D}} R_F + \left(\max_{R_P \,|\, (R_P, R_F) \in \mathcal{AR}(N_P, N_F)} R_P \right) \\
&= \max_{R_F \,|\, (N_F, R_F) \in \mathcal{D}} R_F + \overline{R}_P(N_F, R_F)
\end{aligned} \tag{C.5}$$

From Proposition 12, we know that $R_P^E + R_F^E = \overline{R}(\widetilde{N}) = \overline{R}(N_P + N_F^E)$. Therefore, the equivalence above implies that we must have $R_P^E = \overline{R}_P(N_F, R_F^E)$.

□

Proof. Proof of Proposition 14. Suppose (R_P, R_F, N_F) is an optimal solution of Problem (4.3).

- If $R_F < R_F^E$: because the optimal equal treatment solution is feasible in Problem (4.10), Profit(R_F) must be not less than Profit(R_F^E). Since Profit$(R_F) = R_P + (1-\gamma)R_F - C_P N_P$,

$$\begin{aligned}
& R_P + (1-\gamma)R_F - C_P N_P \geqslant R_P^E + (1-\gamma)R_F^E - C_P N_P \\
\implies & R_P + (1-\gamma)R_F \geqslant R_P^E + (1-\gamma)R_F^E \\
\implies & R_P - R_P^E \geqslant (1-\gamma)(R_F^E - R_F) > 0 \quad \text{since } \gamma \in (0,1) \text{ and } R_F < R_F^E
\end{aligned}$$

this implies $R_P > R_P^E$ and $N_P > 0$.

By Proposition 12, $R_P^E = rN_P/\gamma$ if $N_P < \widetilde{N}$, and $R_P^E = \overline{R}(N_P)$ if $N_P \geqslant \widetilde{N}$. This means $R_P > rN_P/\gamma$ or $R_P > \overline{R}(N_P)$, but the latter is impossible by the definition of $\overline{R}(\cdot)$. Thus, we must have $R_P > rN_P/\gamma$. And because $N_F = \gamma R_F/r$, we either have $N_F = R_F = 0$, or $N_F > 0$ and $R_F/N_F = r/\gamma$. Therefore, (R_P, R_F, N_F) must be a private supply prioritization solution.

- If $R_F > R_F^E$: then $N_F > N_F^E \geqslant 0$ and $N_F + N_P > N_P + N_F^E = \widetilde{N}$.

  For the sake of contradiction, suppose $R_P N_F \geqslant R_F N_P$. Because of the equilibrium,

  $$\overline{R}(N_P + N_F)/(N_P + N_F) \geqslant (R_P + R_F)/(N_P + N_F) \geqslant r/\gamma$$

  Since $\overline{R}(N)$ is continuous, and $\exists M > 0, \lim_{N\to\infty} \overline{R}(N) \leqslant M$, by the intermediate value theorem, there exists $N' \geqslant (N_P + N_F) > (N_P + N_F^E) = \widetilde{N}$ such that $\overline{R}(N')/N' = r/\gamma$. However, this contradicts the definition of $\widetilde{N}$. Thus, we must have $R_P N_F < R_F N_P$. And this implies that (R_P, R_F, N_F) is a flexible supply prioritization solution.

- If $R_F = R_F^E$: then $N_F = N_F^E = \gamma R_F^E/r$. And, by Lemma 26, we have $\overline{R}_P(N_F^E, R_F^E) = R_P^E$. Thus, Proposition 13 implies that the optimal solution must satisfy $R_F = R_F^E$, $N_F = N_F^E$ and $R_P = R_P^E$, which is exactly the optimal equal treatment solution.

□

Proof. Proof of Theorem 8. **Step 1.** [**differentiating the profit**]:

Because $\overline{R}_P$ is differentiable at (N_F^E, R_F^E), we can compute the following derivative:

$$\begin{aligned}
\frac{d\text{profit}}{dR_F}(R_F^E) &= \frac{d\left(R_F \to \overline{R}_P(\gamma R_F/r, R_F)\right)}{dR_F}(R_F^E) + (1-\gamma) \\
&= \frac{\gamma}{r}\frac{\partial \overline{R}_P}{\partial N_F}(\gamma R_F^E/r, R_F^E) + \frac{\partial \overline{R}_P}{\partial R_F}(\gamma R_F^E/r, R_F^E) + (1-\gamma) \\
&= \frac{\gamma}{r}\frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E) + \frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) + (1-\gamma) \quad \text{since } \gamma R_F^E = r N_F^E
\end{aligned}$$

where $\text{profit}(R_F)$ is defined in Proposition 13.

Step 2. [**Small Prioritization is without loss of revenue**]:

Because $N_F^E > 0$ by Assumption 3, we have $N_P < \widetilde{N}$. Proposition 12 gives us $R_P^E + R_F^E = \overline{R}(N_P + N_F^E) = \overline{R}(\widetilde{N})$. And using the optimization problem equivalency at the end of the proof of Lemma 26 (i.e. Equation (C.5)), we have:

$$R_F^E \in \underset{R_F \,|\, (N_F^E, R_F) \in \mathcal{D}}{\operatorname{argmax}} \; R_F + \overline{R}_P(N_F^E, R_F)$$

Therefore, as (N_F^E, R_F^E) is in the interior of $\mathcal{D}$ and as $\overline{R}_P(\cdot,\cdot)$ is differentiable on this point, the derivative in R_F of the above optimization problem must be 0 on $R_F = R_F^E$:

$$\begin{aligned} & \frac{d\left(R_F \to R_F + \overline{R}_P(N_F^E, R_F)\right)}{dR_F}(R_F^E) = 0 \\ \iff \quad & 1 + \frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) = 0 \end{aligned}$$

Therefore, we have $\frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) = -1$. Intuitively, this means that with constant total supply, we can slightly prioritize private agents or flexible agents without modifying the total revenue. Indeed, adding one dollar of revenue to flexible agents would remove one dollar of revenue to private agents.

Step 3. [an interpretable expression for $\frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E)$]:

We define the following function $\gamma(\cdot)$:

$$\gamma(N_F) = \frac{r \cdot (N_P + N_F)}{\overline{R}(N_P + N_F)}$$

Intuitively, $\gamma(N_F)$ represents the commission rate that would be needed to have N_F flexibly supply hours in a revenue-maximizing (equal treatment) solution. Indeed, $\overline{R}(N_P + N_F)$ is the maximum revenue achievable with N_F flexibly supply hours, and we have the flexible supply equilibrium $\gamma(N_F)\overline{R}(N_P + N_F) = r(N_P + N_F)$.

As $\overline{R}(\cdot)$ is differentiable in $\widetilde{N}$, $\gamma(\cdot)$ is also differentiable at $N_F = N_F^E$ (Recall $\widetilde{N} = N_P + N_F^E$ by Proposition 12). We also have $\gamma(N_F^E) = \gamma$, as $\overline{R}(\widetilde{N}) = \frac{r}{\gamma}\widetilde{N}$.

Following the identity (C.5) from the proof of Lemma 26, we know that if $R_F(N_F)$ is the revenue of flexible supply corresponding to the total revenue $\overline{R}(N_P + N_F)$, then we must have:

$$\overline{R}(N_P + N_F) = R_F(N_F) + \overline{R}_P(N_F, R_F(N_F)) \tag{C.6}$$

That is, the corresponding private agent revenue will be given by $\overline{R}_P(N_F, R_F(N_F))$.

Now, using Assumption 2, we know that $\overline{R}(N_P + N_F)$ is always achievable with an equal treatment policy for any N_F, which means that for any $N_F > 0$, there exists $R_F(N_F)$ such

that:

$$\begin{aligned}\frac{R_F(N_F)}{N_F} &= \frac{\overline{R}_P(N_F, R_F(N_F))}{N_P} \\ &= \frac{R_F(N_F) + \overline{R}_P(N_F, R_F(N_F))}{N_F + N_P} \\ &= \frac{\overline{R}(N_P + N_F)}{N_F + N_P} \\ &= \frac{r}{\gamma(N_F)}\end{aligned}$$

Therefore, we can use the identity $R_F(N_F) = rN_F/\gamma(N_F)$ in Equation (C.6) and we obtain:

$$\overline{R}(N_P + N_F) = \frac{rN_F}{\gamma(N_F)} + \overline{R}_P\left(N_F, \frac{rN_F}{\gamma(N_F)}\right)$$

Therefore, we can take the derivative with respect to N_F and we obtain:

$$\begin{aligned}\frac{d\overline{R}}{dN}(\widetilde{N}) =& \frac{d\overline{R}}{dN}(N_P + N_F^E) \\ &= r \cdot \frac{d\left(N_F \to \frac{N_F}{\gamma(N_F)}\right)}{dN_F}(N_F^E) + \frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E) + r \cdot \frac{d\left(N_F \to \frac{N_F}{\gamma(N_F)}\right)}{dN_F}(N_F^E)\frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) \\ &= \frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E) \quad \text{because } \frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) = -1 \text{ by step 2.}\end{aligned}$$

Step 4. [Combining the previous results]

Using the previous results, we obtain immediately:

$$\begin{aligned}\frac{d\text{profit}}{dR_F}(R_F^E) &= \frac{\gamma}{r}\frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E) + \frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) + (1-\gamma) \\ &= \frac{\gamma}{r}\frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E) - 1 + (1-\gamma) \quad \text{since } \frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) = -1 \text{ by step 2.} \\ &= \frac{\gamma}{r}\frac{d\overline{R}}{dN}(\widetilde{N}) - \gamma \quad \text{since } \frac{d\overline{R}}{dN}(\widetilde{N}) = \frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E) \text{ by step 3.} \\ &= \gamma\left(\frac{1}{r}\frac{d\overline{R}}{dN}(\widetilde{N}) - 1\right)\end{aligned}$$

Step 5. [Prioritization]

- Suppose that $\frac{d\overline{R}}{dN}(\widetilde{N}) < r$ and therefore we implement a policy with $R_F = R_F^E + dR_F$ with $dR_F < 0$ to increase profit. We also use the notation $N_F + dN_F$ to denote the

new number of available hours of flexible supply, and $R_P^E + dR_P$ for the new private agent revenue. Then, we can apply Definition 2 and the flexible supply equilibrium, so private agents are prioritized if and only if $N_P r/\gamma < R_P^E + dR_P$. And because $R_P^E = N_P r/\gamma$

$$N_P \frac{r}{\gamma} < R_P^E + dR_P \iff dR_P > 0$$

Note that the profit increase is $dR_P + (1-\gamma)dR_F$, therefore if $dR_F < 0$ and the profit is increased, we must have $dR_P > 0$, which implies that private agents are prioritized.

- Now suppose that $\frac{d\overline{R}}{dN}(\widetilde{N}) > r$ and the firm increases R_F to increase profit, i.e. $R_F = R_F^E + dR_F$ with $dR_F > 0$. Following the same reasoning as the previous point, we need to prove that $dR_P < 0$ to show that flexible agents are prioritized.

  For the sake of contradiction, suppose $dR_P \geqslant 0$. Then, since $R_P^E = N_P r/\gamma$,

  $$R_P^E + dR_P \geqslant N_P \frac{r}{\gamma}$$

  And because of the flexible supply equilibrium, $R_F^E + dR_F = r(N_F^E + dN_F)/\gamma$. Therefore,

  $$\begin{aligned} &\overline{R}(N_P + N_F^E + dN_F) \geqslant R_P^E + dR_P + R_F^E + dR_F \geqslant \frac{r}{\gamma}(N_P + N_F^E + dN_F) \\ \implies &\frac{\overline{R}(N_P + N_F^E + dN_F)}{N_P + N_F^E + dN_F} \geqslant \frac{r}{\gamma} \\ \implies &\frac{\overline{R}(\widetilde{N} + dN_F)}{\widetilde{N} + dN_F} \geqslant \frac{r}{\gamma} \quad \text{since } N_P + N_F^E = \widetilde{N} \end{aligned}$$

  To conclude, note that $N \to \overline{R}(N)/N$ is continuous and goes to 0 when $N \to \infty$. Thus, by the intermediate value theorem, there exists $N' \geqslant \widetilde{N} + dN_F > \widetilde{N}$ such that $\overline{R}(N') = rN'/\gamma$. This contradicts the definition of $\widetilde{N}$. Hence, we must have $dR_P < 0$ and therefore flexible agents are prioritized.

□

Proof. Proof of Theorem 9. Let $\Delta N \geqslant -\widetilde{N}$, because $\overline{R}(N)$ is strictly concave, we have

$$\overline{R}(\widetilde{N} + \Delta N) < \overline{R}(\widetilde{N}) + \overline{R}'(\widetilde{N})\Delta N$$

For any $\Delta N > 0$, if $\overline{R}'(\widetilde{N}) \leqslant r$, we have

$$\overline{R}(\widetilde{N} + \Delta N) - \overline{R}(\widetilde{N}) < \overline{R}'(\widetilde{N})\Delta N \leqslant r\Delta N \tag{C.7}$$

This means if we increase the supply from $\widetilde{N}$ to $\widetilde{N} + \Delta N$, the addition in the maximum revenue is less than $r\Delta N$.

Similarly, for any $-\widetilde{N} \leqslant \Delta N < 0$, if $\overline{R}'(\widetilde{N}) \geqslant r$,

$$\overline{R}(\widetilde{N}) - \overline{R}(\widetilde{N} + \Delta N) > -\overline{R}'(\widetilde{N})\Delta N \geqslant -r\Delta N \tag{C.8}$$

This means if we decrease the supply from $\widetilde{N}$ to $\widetilde{N} + \Delta N$, the loss in the maximum revenue is more than $-r\Delta N$.

Now, consider $(R_P, R_F, N_F^E + \Delta N)$ as an optimal solution of Problem (4.3). Compared with the optimal equal treatment solution, the difference in profit should be

$$\Delta\text{Profit} = R_P - R_P^E + (1-\gamma)(R_F - R_F^E) = R_P + R_F - (R_P^E + R_F^E) - r\Delta N$$

Because we assume $N_F^E > 0$ by Assumption 3 and Proposition 12 shows $R_P^E + R_F^E = \overline{R}(\widetilde{N})$, then we have

$$\Delta\text{Profit} = R_P + R_F - \overline{R}(\widetilde{N}) - r\Delta N \leqslant \overline{R}(\widetilde{N} + \Delta N) - \overline{R}(\widetilde{N}) - r\Delta N$$

Combined with the conclusion in the beginning, if $\overline{R}'(\widetilde{N}) \leqslant r$ and $\Delta N > 0$, then $\Delta\text{Profit} < 0$ by Inequality (C.7). And if $\overline{R}'(\widetilde{N}) \geqslant r$ and $-\widetilde{N} \leqslant \Delta N < 0$, then $\Delta\text{Profit} < 0$ by Inequality (C.8). Since we assume $(R_P, R_F, N_F^E + \Delta N)$ is optimal, this means that if $\overline{R}'(\widetilde{N}) < r$, we must have $\Delta N \leqslant 0$; and if $\overline{R}'(\widetilde{N}) > r$, we must have $\Delta N \geqslant 0$; and if $\overline{R}'(\widetilde{N}) = r$, we must have $\Delta N = 0$.

Hence, if $\overline{R}'(\widetilde{N}) = r$, then $\Delta N = 0$ and $R_F = R_F^E$. And by Proposition 14, any optimal policy must be an equal treatment policy.

If $\overline{R}'(\widetilde{N}) < r$, then $\Delta N \leqslant 0$. And by Theorem 8, there exists a private supply prioritization policy that has a higher profit than any equal treatment solution, so we must have

$\Delta N < 0$ and $R_F < R_F^E$. Thus, by Proposition 14, any optimal policy must be a private supply prioritization policy.

If $\overline{R}'(\widetilde{N}) > r$, then $\Delta N \geqslant 0$. And by Theorem 8, there exists a flexible supply prioritization policy that has a higher profit than any equal treatment solution, so we must have $\Delta N > 0$ and $R_F > R_F^E$. Thus, by Proposition 14, any optimal policy must be a flexible supply prioritization policy.

□

Proof. Proof of Proposition 15. As shown in Lemma 5, $(R_P^\star, R_F^\star, N_F^\star)$ is also an optimal solution of Problem (4.7). And in Problem (4.7), we can see that given $N_F^\star$, it is optimal to maximize the revenue, so $\widetilde{N}^\star = N_P + N_F^\star$ must be also an optimal solution of

$$\max_{N \geqslant N_P} \overline{R}(N) - r(N - N_P)$$

where we change the variable $N = N_P + N_F$.

Suppose that $N_F^\star > 0$, and $\overline{R}(\cdot)$ is strictly concave. Therefore, 0 must belong to the subderivative of this objective at $N = \widetilde{N}^\star$, and the derivative of the objective at $N = \widetilde{N}$ is $\frac{d\overline{R}}{dN}(\widetilde{N}) - r$. Thus, the strict concavity implies that:

$$N_F^\star < N_F^E \iff \frac{d\overline{R}}{dN}(\widetilde{N}) < r$$

and

$$N_F^\star > N_F^E \iff \frac{d\overline{R}}{dN}(\widetilde{N}) > r$$

Notice that $\frac{d\overline{R}}{dN}(\widetilde{N}) = \frac{d\overline{R}}{dN}(N_P + N_F^E)$ by Assumption 3 and Proposition 12.

In addition, if $N_F^\star = 0$, it means any increase in N from N_P will lead to a lower profit, so we must have $N_F^\star < N_F^E$ and

$$\frac{d\overline{R}}{dN}(N_P + N_F^E) = \frac{d\overline{R}}{dN}(\widetilde{N}) < r$$

Moreover, in the proof Theorem 8 and Theorem 9, we have seen that if $\overline{R}'(\widetilde{N}) < r$, then the profit can be improved by a negative deviation of total supply (i.e. $\Delta N < 0$) from $\widetilde{N}$;

and if $\overline{R}'(\widetilde{N}) > r$, then the profit can be improved a positive deviation of total supply (i.e. $\Delta N > 0$) from $\widetilde{N}$.

□

In the following, we look into the consequences of the prioritization on revenue. To this end, we introduce the notation

$$\text{revenue}(R_F) \triangleq \overline{R}_P(\gamma R_F/r, R_F) + R_F,$$

which corresponds to the revenue of the profit-maximizing policy given R_F. We combine Proposition 19 and Theorem 8 to determine the effect of prioritization.

Proposition 19 (Change in Revenue)**.** We have,

$$\frac{d\text{revenue}}{dR_F}(R_F^E) = \frac{\gamma}{r}\overline{R}'(\widetilde{N}).$$

And therefore,

$$\frac{d\text{revenue}}{d\text{Profit}}(R_F^E) = \left(1 - \frac{r}{\overline{R}'(\widetilde{N})}\right)^{-1}.$$

Proof. Proof of Proposition 19. Because $\text{revenue}(R_F) = \overline{R}_P(\gamma R_F/r, R_F) + R_F$, we have $\text{revenue}(R_F) = \text{profit}(R_F) + \gamma R_F$. Thus, by step 1 in the proof of Theorem 8, we can obtain:

$$\frac{d\text{revenue}}{dR_F}(R_F^E) = \frac{\gamma}{r}\frac{\partial \overline{R}_P}{\partial N_F}(N_F^E, R_F^E) + \frac{\partial \overline{R}_P}{\partial R_F}(N_F^E, R_F^E) + 1$$

And from step 4 in the proof of Theorem 8, we can get:

$$\frac{d\text{revenue}}{dR_F}(R_F^E) = \frac{\gamma}{r}\frac{d\overline{R}}{dN}(\widetilde{N})$$

Thus, we have:

$$\frac{d\text{revenue}}{d\text{profit}}(R_F^E) = \frac{\frac{d\text{revenue}}{dR_F}(R_F^E)}{\frac{d\text{profit}}{dR_F}(R_F^E)} = \frac{1}{1 - \frac{r}{\frac{d\overline{R}}{dN}(\widetilde{N})}}$$

□

C.4 Proofs for Section 4.5.

Proof. Proof of Proposition 16 Notice that in the proof of Corollary 2, we show that for any γ, if the optimal solution has $N_F > 0$, we can always find another feasible solution that produces a higher objective value with $N_F = 0$. Thus, the proof here is exactly the same with the proof in Corollary 2.

□

To prove Theorem 10, we need the following auxiliary lemma. Given α and N_P, let $N_F^\alpha(N_P)$ to denote the optimal number of available hours of flexible supply that maximizes the profit in equilibrium. In other words, $N_F^\alpha(N_P)$ is the optimal solution of the following Problem (C.9):

$$\begin{aligned} \max_{R_F, N_F} \quad & (1-\gamma)R_F \\ \text{s.t.} \quad & rN_F = \gamma R_F, \\ & R_P = \alpha r N_P/\gamma \\ & (R_P, R_F) \in \mathcal{AR}(N_P, N_F). \end{aligned} \tag{C.9}$$

Notice that $N_F^\alpha(0) = \widetilde{N}$.

Lemma 27 (Change of Flexible Supply)**.** Suppose Assumption 1 holds. Given a well-defined α and β^α at $N_P = 0$, the gradient of the optimal solution of Problem (C.9) with respect to N_P can be expressed as:

$$\frac{\partial N_F^\alpha}{\partial N_P}(0) = -\left(1 + \frac{(\alpha-1) + \gamma\beta_0^\alpha}{1 - \gamma\overline{R}'(\widetilde{N})/r}\right) \tag{C.10}$$

Proof. Proof of Lemma 27 Given a well-defined α verifying Definition 4, the maximum revenue which flexible supply may receive is:

$$\overline{R}(N_P + N_F) - \alpha r N_P/\gamma - \beta^\alpha(N_P, N_F) r N_P$$

With the flexible supply equilibrium, we define:

$$F^{\alpha}(N_P, N_F) \triangleq \overline{R}(N_P + N_F) - \alpha r N_P/\gamma - \beta^{\alpha}(N_P, N_F) r N_P - r N_F/\gamma$$

Basically, $F^{\alpha}(N_P, N_F)$ is the difference between the maximum revenue of flexibly supply and the in-equilibrium revenue of flexibly supply, given N_P, N_F and α. Now, in order to use the implicit differentiation and derive $\frac{\partial N_F^{\alpha}}{\partial N_P}$ at $N_P = 0$, we need to show $N_F^{\alpha}(N_P)$ satisfies $F^{\alpha}(N_P, N_F^{\alpha}(N_P)) = 0$ in a neighborhood of $N_P = 0$. First, by Definition 4, let $U \times V$ denote a neighborhood of point $(0, \widetilde{N})$ in which $\beta^{\alpha}(N_P, N_F)$ is continuous, so $F^{\alpha}(N_P, N_F)$ is also continuous in $U \times V$. Then, the rest of the proof is in the following steps.

Step 1. $[\forall N > \widetilde{N}, F^{\alpha}(0, N) < 0]$

By definition, $\beta^{\alpha}(N_P, N) r N_P = \Delta R^{\alpha}(N_P, N_F)$ and $\Delta R^{\alpha}(0, N) = 0$ by Assumption 1, so we have

$$F^{\alpha}(0, N) = \overline{R}(N) - rN/\gamma$$

For the sake of contradiction, if $N > \widetilde{N}$ but $F^{\alpha}(0, N) \geqslant 0$, it means $\overline{R}(N) - rN/\gamma \geqslant 0$ so that $\overline{R}(N)/N \geqslant r/\gamma$. However, because $\overline{R}(N)$ is continuous and $\lim_{N\to\infty} \overline{R}(N) < \infty$, there would exist $N' \geqslant N > \widetilde{N}$ such that $\overline{R}(N)/N = r/\gamma$. This contradicts the definition of $\widetilde{N}$. Therefore, we must have $\forall N > \widetilde{N}, F^{\alpha}(0, N) < 0$.

Step 2. $[\exists N < \widetilde{N}$ **and** $N \in V, F^{\alpha}(0, N) > 0]$

Since we assume $\overline{R}'(\widetilde{N}) < r/\gamma$, by the definition of derivative, we can pick an $\epsilon \in (0, r/\gamma - \overline{R}'(\widetilde{N}))$, there exists $\delta > 0$ such that

$$-\delta < h < 0 \implies \frac{\overline{R}(\widetilde{N} + h) - \overline{R}(\widetilde{N})}{h} - \overline{R}'(\widetilde{N}) < \epsilon$$

This implies

$$\begin{aligned}
&\frac{\overline{R}(\widetilde{N}+h)-\overline{R}(\widetilde{N})}{h} < \epsilon + \overline{R}'(\widetilde{N}) \\
\Longrightarrow\ &\frac{\overline{R}(\widetilde{N}+h)-\overline{R}(\widetilde{N})}{h} < \frac{r}{\gamma} \quad \text{because } \epsilon < r/\gamma - \overline{R}'(\widetilde{N}) \\
\Longrightarrow\ &\overline{R}(\widetilde{N}+h)-\overline{R}(\widetilde{N}) > \frac{r}{\gamma}h \quad \text{because } h<0 \\
\Longrightarrow\ &\overline{R}(\widetilde{N}+h) > \frac{r}{\gamma}(\widetilde{N}+h) \quad \text{because } \gamma\overline{R}(\widetilde{N}) = r\widetilde{N} \\
\Longrightarrow\ &\overline{R}(N) > \frac{r}{\gamma}N \quad \text{let } N = \widetilde{N}+h
\end{aligned}$$

Hence, we can choose $N \in (\widetilde{N}-\delta, \widetilde{N}) \cap V$ such that $\overline{R}(N) > rN/\gamma$ and therefore, $F^{\alpha}(0,N) > 0$.

Step 3. [Construct a neighborhood of $N_P = 0$ such that $F^{\alpha}(N_P, N_F^{\alpha}(N_P)) = 0$]

In the following, we want to construct a neighborhood $\widetilde{U} \times \widetilde{V} \subset U \times V$ of point $(0, \widetilde{N})$ such that for any $N_P \in \widetilde{U}$, $N_F^{\alpha}(N_P)$ must reside in $\widetilde{V}$ and justify $F^{\alpha}(N_P, N_F^{\alpha}(N_P)) = 0$ by the intermediate value theorem.

1. Because $\widetilde{N}$ is an interior point of V, there exists $\xi > 0$ such that $(0, \widetilde{N}+\xi) \in U \times V$. By Step 1, we know $F^{\alpha}(0, \widetilde{N}+\xi) < 0$. Because $F^{\alpha}(N_P, N_F)$ is continuous at point $(0, \widetilde{N}+\xi)$, then $\forall \epsilon > 0$, there exists $\delta > 0$ such that $N_P < \delta \implies |F^{\alpha}(N_P, \widetilde{N}+\xi) - F^{\alpha}(0, \widetilde{N}+\xi)| < \epsilon$. Pick an $\epsilon \in (0, -F^{\alpha}(0, \widetilde{N}+\xi))$, there exists $\delta > 0$ such that

$$N_P < \delta \implies F^{\alpha}(0, \widetilde{N}+\xi) - \epsilon < F^{\alpha}(N_P, \widetilde{N}+\xi) < F^{\alpha}(0, \widetilde{N}+\xi) + \epsilon < 0$$

   Notice that $F^{\alpha}(0, \widetilde{N}+\xi) + \epsilon < 0$ is due to $\epsilon < -F^{\alpha}(0, \widetilde{N}+\xi)$. Thus, this means

$$N_P < \delta \implies F^{\alpha}(N_P, \widetilde{N}+\xi) < 0$$

2. By Step 2, we can choose a $\nu > 0$ such that $\widetilde{N} - \nu \in V$ and $F^{\alpha}(0, \widetilde{N}-\nu) > 0$. Then, we can repeat a similar logic as the above. Because $F^{\alpha}(N_P, N_F)$ is continuous at point $(0, \widetilde{N}-\nu)$, then $\forall \epsilon > 0$, there exists $\theta > 0$ such that $N_P < \theta \implies |F^{\alpha}(N_P, \widetilde{N}-\nu) -$

$F^\alpha(0, \widetilde{N} - \nu)| < \epsilon$. Pick an $\epsilon \in (0, F^\alpha(0, \widetilde{N} - \nu))$, there exists $\theta > 0$ such that

$$N_P < \theta \implies 0 < F^\alpha(0, \widetilde{N} - \nu) - \epsilon < F^\alpha(N_P, \widetilde{N} - \nu) < F^\alpha(0, \widetilde{N} - \nu) + \epsilon$$

Notice that $F^\alpha(0, \widetilde{N} - \nu) - \epsilon > 0$ is due to $\epsilon < F^\alpha(0, \widetilde{N} - \nu)$. Thus, this means

$$N_P < \theta \implies F^\alpha(N_P, \widetilde{N} - \nu) > 0$$

3. Let $\mu = \sup_{N \geqslant \widetilde{N} + \xi} F^\alpha(0, N)$. We first show $\mu < 0$. Because we assumed $\exists M > 0$ such that $\lim_{N \to \infty} \overline{R}(N) < M$, this means $F^\alpha(0, N) \to -\infty$ as $N \to \infty$. That is, for $K < 0$, there exists $N' > 0$ such that $\forall N > N'$, $F^\alpha(0, N) < K$. If $N' < \widetilde{N} + \xi$, then $\mu \leqslant K < 0$. If $N' \geqslant \widetilde{N} + \xi$, then $[\widetilde{N} + \xi, N']$ is compact, and we know $F^\alpha(0, N)$ is continuous, so there exists $k \in [\widetilde{N} + \xi, N']$ such that $\forall N \in [\widetilde{N} + \xi, N']$, $F^\alpha(0, N) \leqslant F^\alpha(0, k)$. And because $\forall N > \widetilde{N}$, $F^\alpha(0, N) < 0$ by Step 1, $F^\alpha(0, k) < 0$. This implies $\mu \leqslant \max\{K, F^\alpha(0, k)\} < 0$. Therefore, $\mu < 0$.

   Then we want to show $\overline{R}(N)$ is uniformly continuous. Let $L \triangleq \lim_{N \to \infty} \overline{R}(N)$ which is finite by our assumption. For any $\epsilon > 0$, there exists $N' > 0$ such that $\forall N > N'$, $|\overline{R}(N) - L| < \epsilon/2$. Because $\overline{R}(N)$ is continuous and $[0, N' + 1]$ is compact, $\overline{R}(N)$ is uniformly continuous on $[0, N' + 1]$. This implies that there exists $\delta > 0$ such that $\forall n, m \in [0, N' + 1]$, $|n - m| < \min\{\delta, 1\} \implies |\overline{R}(n) - \overline{R}(m)| < \epsilon$. Now, for any $n, m \geqslant 0$, $|n - m| < \min\{\delta, 1\}$, they are either in $[0, N' + 1]$ or (N', ∞), if $n, m \in [0, N' + 1]$, we know $|\overline{R}(n) - \overline{R}(m)| < \epsilon$; whereas if $n, m \in (N', \infty)$, we have $|\overline{R}(n) - \overline{R}(m)| \leqslant |\overline{R}(n) - L| + |L - \overline{R}(m)| < \epsilon/2 + \epsilon/2 = \epsilon$. Therefore, $\overline{R}(N)$ is uniformly continuous on $[0, \infty)$.

   The uniform continuity of $\overline{R}(N)$ implies that we can choose $\phi \in (0, -\mu)$, there exists $\psi > 0$ such that $N_P < \psi \implies \forall N_F \geqslant 0$, $|\overline{R}(N_F + N_P) - \overline{R}(N_F)| < \phi$. Therefore,

$$\begin{aligned} F^\alpha(N_P, N_F) - F^\alpha(0, N_F) &= \overline{R}(N_P + N_F) - \alpha r N_P/\gamma - \beta^\alpha(N_P, N_F) r N_P - \overline{R}(N_F) \\ &\leqslant \overline{R}(N_P + N_F) - \overline{R}(N_F) < \phi \end{aligned}$$

This follows that $N_P < \psi \implies \forall N \geqslant \widetilde{N}+\xi,\ F^\alpha(N_P, N) < F^\alpha(0, N)+\phi$. And because $\phi \in (0, -\mu)$, we finally get

$$N_P < \psi \implies \forall N \geqslant \widetilde{N} + \xi,\ F^\alpha(N_P, N) < 0$$

4. Let $\omega \triangleq \min\{\delta, \theta, \psi\}$, $\widetilde{U} \triangleq [0, \omega) \cap U$ and $\widetilde{V} \triangleq [\widetilde{N} - \nu, \widetilde{N} + \xi] \cap V$. And because $\widetilde{N} - \nu, \widetilde{N} + \xi \in V$, the final neighborhood of $(0, \widetilde{N})$ that we want to construct is

$$\widetilde{U} = [0, \omega) \cap U,\ \widetilde{V} = [\widetilde{N} - \nu, \widetilde{N} + \xi]$$

Notice that $(0, \widetilde{N}) \in \widetilde{U} \times \widetilde{V}$.

Step 4. [Verify $F^\alpha(N_P, N_F^\alpha(N_P)) = 0$ for $N_P \in \widetilde{U}$.]

Let $N_P \in \widetilde{U}$, we first want to show there exists $N_F' \in \widetilde{V}$ such that $F^\alpha(N_P, N_F') = 0$. It follows showing $N_F^\alpha(N_P) \in \widetilde{V}$. And we will finally get $F^\alpha(N_P, N_F^\alpha(N_P)) = 0$.

First, because $F^\alpha(N_P, \widetilde{N} + \xi) < 0$, $F^\alpha(N_P, \widetilde{N} - \nu) > 0$, and $F^\alpha(N_P, N_F)$ is continuous in $\widetilde{U} \times \widetilde{V}$, by the intermediate value theorem, there exists $N_F' \in \widetilde{V}$ such that $F^\alpha(N_P, N_F') = 0$. This means N_F' is a feasible solution of Problem (C.9) and $N_F' \leqslant N_F^\alpha(N_P)$.

Second, to see $N_F^\alpha(N_P) \in \widetilde{V}$, we first notice that any $N < \widetilde{N} - \nu$ cannot be $N_F^\alpha(N_P)$, because we have shown there exists a feasible solution for $N \in \widetilde{V}$ in the above. Therefore, we only need to check whether the point $N_F > \widetilde{N} + \xi$ may be feasible in Problem (C.9). In fact, because $\forall N \geqslant \widetilde{N} + \xi,\ F^\alpha(N_P, N) < 0$, it means that the maximum revenue of flexible supply is always less than the equilibrium. That is, if $N \geqslant \widetilde{N}+\xi$, then $\forall R_F, (\alpha r N_P/\gamma, R_F) \in \mathcal{AR}(N_P, N), R_F < rN/\gamma$. Thus, any $N_F > \widetilde{N} + \xi$ cannot be feasible in Problem (C.9).

The above two points indicate $N_F^\alpha(N_P) \in \widetilde{V}$. Now, for the sake of contradiction, suppose $F^\alpha(N_P, N_F^\alpha(N_P)) \neq 0$. Since we know there exists an achievable R_F such that $R_F = rN_F^\alpha(N_P)/\gamma$, it implies $F^\alpha(N_P, N_F^\alpha(N_P)) > 0$. Because $F^\alpha(N_P, \widetilde{N} + \xi) < 0$ and $F^\alpha(N_P, N_F)$ is continuous in $\widetilde{U} \times \widetilde{V}$, by the intermediate value theorem, there exists $N_F' \in (N_F^\alpha(N_P), \widetilde{N}+\xi)$ such that $F^\alpha(N_P, N_F') = 0$. This means N_F' is feasible in Problem (C.9) and $N_F' > N_F^\alpha(N_P)$, which contradicts the assumption that $N_F^\alpha(N_P)$ is the optimal solution of Problem (C.9).

Hence, we conclude that $F^\alpha(N_P, N_F^\alpha(N_P)) = 0$ for $N_P \in \widetilde{U}$.

Step 5. [Compute $\frac{\partial N_F^\alpha}{\partial N_P}(N_P = 0)$.]

Now we are able to apply the implicit differentiation, by Definition 4, we can compute the partial derivative of $F^\alpha(N_P, N_F)$ at $(0, \widetilde{N})$:

$$\frac{\partial F^\alpha}{\partial N_P}(N_P, N_F) = \overline{R}'(N_P + N_F) - \alpha\frac{r}{\gamma} - \beta^\alpha(N_P, N_F)r - \frac{\partial \beta^\alpha}{\partial N_P}(N_P, N_F)rN_P$$
$$\frac{\partial F^\alpha}{\partial N_F}(N_P, N_F) = \overline{R}'(N_P + N_F) - \frac{r}{\gamma} - \frac{\partial \beta^\alpha}{\partial N_F}(N_P, N_F)rN_P$$

At $N_P = 0$, because $N_F^\alpha(0) = \widetilde{N}$ for any feasible α,

$$\frac{\partial F^\alpha}{\partial N_P}(0, N_F^\alpha(0)) = \frac{\partial F^\alpha}{\partial N_P}(0, \widetilde{N}) = \overline{R}'(\widetilde{N}) - \alpha\frac{r}{\gamma} - \beta_0^\alpha r$$
$$\frac{\partial F^\alpha}{\partial N_F}(0, N_F^\alpha(0)) = \frac{\partial F^\alpha}{\partial N_F}(0, \widetilde{N}) = \overline{R}'(\widetilde{N}) - \frac{r}{\gamma}$$

Notice that $\frac{\partial F^\alpha}{\partial N_F}(0, \widetilde{N}) \neq 0$, since $\overline{R}'(\widetilde{N}) \neq r/\gamma$ by Definition 4.

Then, the derivative $\frac{\partial N_F^\alpha}{\partial N_P}(0)$ can be computed as:

$$\frac{\partial N_F^\alpha}{\partial N_P}(0) = -\frac{\partial F^\alpha / \partial N_P}{\partial F^\alpha / \partial N_F}(0, \widetilde{N})$$

Hence, we can finally get:

$$\frac{\partial N_F^\alpha}{\partial N_P}(0) = \frac{\alpha r/\gamma + \beta_0^\alpha r - \overline{R}'(\widetilde{N})}{\overline{R}'(\widetilde{N}) - r/\gamma} = \frac{(\alpha - 1)r/\gamma + r\beta_0^\alpha}{\overline{R}'(\widetilde{N}) - r/\gamma} - 1 = -\left(1 + \frac{(\alpha - 1) + \gamma\beta_0^\alpha}{1 - \gamma\overline{R}'(\widetilde{N})/r}\right)$$

□

Proof. Proof of Theorem 10 We know the profit is $R_P + (1 - \gamma)R_F - C_P N_P$. Given N_P and α, $R_P = \alpha r N_P/\gamma$, and with the flexible supply equilibrium, the optimal R_F is equal to $rN_F^\alpha(N_P)/\gamma$, where $N_F^\alpha(N_P)$ is defined in Problem (C.9). Therefore, the optimal profit in equilibrium can be expressed as:

$$\text{profit}(N_P) = \alpha\frac{r}{\gamma}N_P + (1 - \gamma)\frac{r}{\gamma}N_F^\alpha(N_P) - C_P N_P$$

By taking the derivative with respect to N_P and applying Lemma 27, we have:

$$\begin{aligned}\frac{d\text{profit}}{dN_P}(N_P=0) &= \alpha\frac{r}{\gamma} - (1-\gamma)\frac{r}{\gamma}\cdot\left(1+\frac{(\alpha-1)+\gamma\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r}\right) - C_P \\ &= \left(\alpha\frac{r}{\gamma} - C_P\right) - (1-\gamma)\frac{r}{\gamma}\cdot\left(1+\frac{(\alpha-1)+\gamma\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r}\right)\end{aligned}$$

Notice that it is optimal to introduce private supply into the market with the level of prioritization α if and only if the above derivative is non-negative,

$$\frac{d\text{profit}}{dN_P}(N_P=0) = \left(\alpha\frac{r}{\gamma} - C_P\right) - (1-\gamma)\frac{r}{\gamma}\cdot\left(1+\frac{(\alpha-1)+\gamma\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r}\right) \geqslant 0$$

Rearrange the above inequality to get Inequality (4.15),:

$$\begin{aligned}&\left(\alpha\frac{r}{\gamma} - C_P\right) - (1-\gamma)\frac{r}{\gamma}\cdot\left(1+\frac{(\alpha-1)+\gamma\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r}\right) \geqslant 0 \\ \iff &\frac{\alpha}{\gamma} - \frac{1-\gamma}{\gamma}\cdot\left(1+\frac{(\alpha-1)+\gamma\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r}\right) \geqslant \frac{C_P}{r} \\ \iff &\frac{\alpha-1}{\gamma} - \frac{1-\gamma}{\gamma}\cdot\frac{(\alpha-1)+\gamma\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r} \geqslant \frac{C_P}{r} - 1 \\ \iff &\frac{(\alpha-1)(1-\gamma\overline{R}'(\widetilde{N})/r) - (1-\gamma)[(\alpha-1)+\gamma\beta_0^\alpha]}{\gamma(1-\gamma\overline{R}'(\widetilde{N})/r)} \geqslant \frac{C_P}{r} - 1 \\ \iff &\frac{(\alpha-1)(1-\gamma\overline{R}'(\widetilde{N})/r-1+\gamma) + (1-\gamma)\gamma\beta_0^\alpha}{\gamma(1-\gamma\overline{R}'(\widetilde{N})/r)} \geqslant \frac{C_P}{r} - 1 \\ \iff &\frac{(\alpha-1)(1-\overline{R}'(\widetilde{N})/r) - (1-\gamma)\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r} \geqslant \frac{C_P}{r} - 1 \\ \iff &\frac{(\alpha-1)(1-\overline{R}'(\widetilde{N})/r)}{1-\gamma\overline{R}'(\widetilde{N})/r} \geqslant \frac{(1-\gamma)\beta_0^\alpha}{1-\gamma\overline{R}'(\widetilde{N})/r} + \frac{C_P}{r} - 1\end{aligned}$$

□

Proof. Proof of Theorem 11 In this proof, we present an example to show the possibility that a prioritization policy with hybrid is optimal when $C_P > r$. Suppose $\overline{R}(\cdot)$ is increasing, strictly concave and differentiable, then the optimal number of available hours in a private-

supply-only market can be given by:

$$N_P^\star = \arg\max_{N_P \geqslant 0} \overline{R}(N_P) - C_P N_P$$

And assume C_P is not larger than $\overline{R}'(0)$, so $N_P^\star$ verifies the first order condition: $\overline{R}'(N_P^\star) = C_P$.

Let $\gamma = \frac{rN_P^\star}{\overline{R}(N_P^\star)}$. Since $\overline{R}(N_P^\star) - C_P N_P^\star \geqslant 0$ and $C_P > r$, we must have $\gamma \in (0, 1)$.

We want to show that in this case, (1) the optimal flexible-supply-only solution has a higher profit than any private-supply-only solution; (2) it is possible that a hybrid-supply solution has a higher profit than the optimal flexible-supply-only solution.

1. Recall that the optimal profit with flexible supply only is given by:

$$\max_{N_F \geqslant 0} \overline{R}(N_F) - rN_F \quad \text{s.t.} \quad \gamma \overline{R}(N_F) = rN_F$$

   Since $\gamma = \frac{rN_P^\star}{\overline{R}(N_P^\star)}$, $N_F = N_P^\star$ is a feasible solution of the above problem with the objective equal to $\overline{R}(N_P^\star) - rN_P^\star$, which is larger than $\overline{R}(N_P^\star) - C_P N_P^\star$. Therefore, the optimal profit with flexible supply only is larger than the optimal profit with private supply only.

2. Now suppose that $\beta_0^\alpha = 0$ for any feasible α (e.g. in a queueing model), by Inequality (4.16), we need

$$(\alpha - 1)\frac{1 - \overline{R}'(\widetilde{N})/r}{1 - \gamma \overline{R}'(\widetilde{N})/r} \geqslant \frac{C_P}{r} - 1$$

   And because $\overline{R}(N)$ is strictly concave and $\overline{R}(N_P^\star)/N_P^\star = r/\gamma$, then for any $N > N_P^\star$,

$\overline{R}(N)/N < r/\gamma$. This implies $\widetilde{N} = N_P^\star$. And we know $\overline{R}'(N_P^\star) = C_P$, so

$$\begin{aligned}
&(\alpha-1)\frac{1-\overline{R}'(\widetilde{N})/r}{1-\gamma\overline{R}'(\widetilde{N})/r} \geqslant \frac{C_P}{r} - 1 \\
\iff &(\alpha-1)\frac{1-\overline{R}'(N_P^\star)/r}{1-\gamma\overline{R}'(N_P^\star)/r} \geqslant \frac{C_P}{r} - 1 \\
\iff &(\alpha-1)\frac{1-C_P/r}{1-\gamma C_P/r} \geqslant \frac{C_P}{r} - 1 \\
\iff &(\alpha-1)(1-C_P/r) \geqslant (\frac{C_P}{r} - 1)(1-\gamma C_P/r)
\end{aligned}$$

Notice that since $\overline{R}(\cdot)$ is strictly concave, $C_P = \overline{R}'(N_P^\star) = \overline{R}'(\widetilde{N}) < \overline{R}(\widetilde{N})/\widetilde{N} = r/\gamma$, which implies $1 - \gamma C_P/r > 0$.

Additionally, because $C_P > r$,

$$(\alpha-1)(1-C_P/r) \geqslant (\frac{C_P}{r} - 1)(1-\gamma C_P)/r) \iff \alpha \leqslant \frac{\gamma C_P}{r}$$

Note that since $C_P < r/\gamma$, we have $\alpha < 1$, which is consistent with what we conclude in Section 4.5

Hence, in this case, a hybrid-supply prioritization policy is optimal if we are able to set the level of prioritization to be less than $\gamma C_P/r$. By the definition of α, this means we need to set the average revenue of private supply to be less than or equal to C_P (i.e. $R_P < C_P N_P$). An example of $\alpha \leqslant C_P$ is to fully prioritize flexible supply such that the revenue of flexible supply is not reduced (i.e. $R_F \geqslant \overline{R}(\widetilde{N})$). Then, the revenue of private supply must satisfy $R_P \leqslant \overline{R}(\widetilde{N} + N_P) - \overline{R}(\widetilde{N})$. Because $\overline{R}'(\widetilde{N}) = C_P$ and $\overline{R}(\cdot)$ is strictly concave, we have

$$\overline{R}(\widetilde{N} + N_P) - \overline{R}(\widetilde{N}) < C_P N_P \implies R_P < C_P N_P$$

To complete the proof, we provide a more concrete example with a queueing model to explain how the above specifications are achievable. Consider a queue of servers waiting for the requests. The arrivals of servers and requests are independent and follow some distribution (e.g. an M/M/1 queque). Once a request arrives, it will be matched with the

first available server in the queue (i.e. on a first-come-first-serve basis). If there are no servers available, the request is lost. Once a server is successfully matched with a request, they will leave the system, and a reward will be given. Therefore, the total maximum revenue (i.e. $\overline{R}(\cdot)$) depends on the departure rate, which further relies on the average number of servers in the system. In addition, because the more servers are in the system, the longer expected waiting time we have, so the marginal revenue is diminishing with respect to the total number of servers. Thus, $\overline{R}(\cdot)$ is increasing, strictly concave, and differentiable in this scenario.

Now, suppose there are two types of servers: private servers and flexible servers, which act as private agents and flexible agents (i.e. there exists C_P, r, γ and the equilibrium constraint in the system). The average number of servers of each type (i.e. N_P and N_F) is constant. The requests are indifferent between the types of servers, and the reward is independent of the type of the matched server, so Assumption 1 and Assumption 2 hold in this case.

For simplicity, the firm is only able to choose to either equally treat two types of the servers or prioritize some servers. If the firm chooses to equally treat them, then the queue is running as the above, whereas if the firm chooses to prioritize some servers, it can move one of the servers in the queue to the first position. The prioritization may be random (e.g. only select a part of private servers randomly), so the level of prioritization relies on how often we prioritize a server. For instance, if we always move all the private servers in the queue to the front of the line, α is maximized, whereas if we always move all the flexible servers in the queue to the front of the line, α is minimized. Because the arrival of the requests is independent of the servers, any prioritization will not affect the total revenue (i.e. $\beta_0^\alpha = 0$).

Hence, we can see that this concrete example satisfies all the assumptions. Moreover, when we fully prioritize the flexible servers, the revenue of the flexible servers will be unaffected by any introduction of the private servers (i.e. $R_F = \overline{R}(\widetilde{N})$), and the private servers will take the remaining revenue (i.e. $R_P = \overline{R}(\widetilde{N} + N_P) - \overline{R}(\widetilde{N})$), so it is feasible to set $R_P/N_P \leqslant C_P$ when $\overline{R}'(\widetilde{N}) = C_P$ as what we discuss above.

□

C.5 A Geometrical View

We provide a geometrical interpretation of the achievable revenues set and how the symmetry and the equal treatment assumptions (Assumption 1 and Assumption 2) shape its structure. This geometrical interpretation will lead us to a natural reformulation of Problem (4.3) in Proposition 13.

To fix ideas, let us consider $\mathcal{AR}(N) = [0, \sqrt{N}]$, that is, in a single-type setting with N supply hours available a firm can garner at most $\sqrt{N}$ total revenue. We now see the implications of this for the geometry of the achievable revenues set $\mathcal{AR}(N_P, N_F)$ in a two-type setting. Consider Figure C.1 (a), when $N_F = 0$ (there is only private supply) the firm's achievable revenues coincide with what can be achieved in a single-type setting with N_P supply hours available. This is shown by the red-thick line in Figure C.1 (a). Consider next the more general case when $N_F > 0$. By our symmetry assumption (c.f., Assumption 1), the achievable revenue pairs must be such that their total revenue coincides with what can be achieved in a single-type system with $N_P + N_F$ supply hours. In the latter system, the achievable total revenue varies from 0 to $\sqrt{N_P + N_F}$. The blue region in Figure C.1 (a), corresponds to the achievable revenues in a two-type system. In accordance with symmetry, this region includes pairs (R_P, R_F) that add up to any total revenue in $[0, \sqrt{N_P + N_F}]$. For example, $\mathcal{AR}(N_P, N_F)$ is tangent to the line $R_P - \sqrt{N_P + N_F}$ (dashed line (1) in the figure), that is, $\sqrt{N_P + N_F}$ is achieved once; additionally, this set intersect at several points with the line $R_P - \sqrt{N_P}$ (dashed line (2) in the figure) so that there are various ways of achieving a total revenue of $\sqrt{N_P}$. Consequently, our symmetry assumption can be interpreted as $\mathcal{AR}(N_P, N_F)$ having non-empty intersection (only) with all lines such that $R_P + R_F \leqslant \sqrt{N_P + N_F}$.

Let us consider now Assumption 2 about equal treatment policies. We depict pairs of

revenue that satisfy equal treatment revenue in Figure C.1 (a) by the black line with slope N_F/N_P (for the case $N_F > 0$). The assumption requires that any revenue in $[0, \sqrt{N_P + N_F}]$ that is achievable in a single-type setting can also be achieved by an equal treatment policy in the two-type setting. In Figure C.1 (a), this means that the intersection of the equal treatment line with any line of the form $R_P + R_F = R$ for $R \leqslant \sqrt{N_P + N_F}$ lies inside the blue region, $\mathcal{AR}(N_P, N_F)$. Additionally, note that a pair (R_P, R_F) that lies below the equal treatment line prioritizes private supply, while a pair that lies above the equal treatment line prioritizes flexible supply.

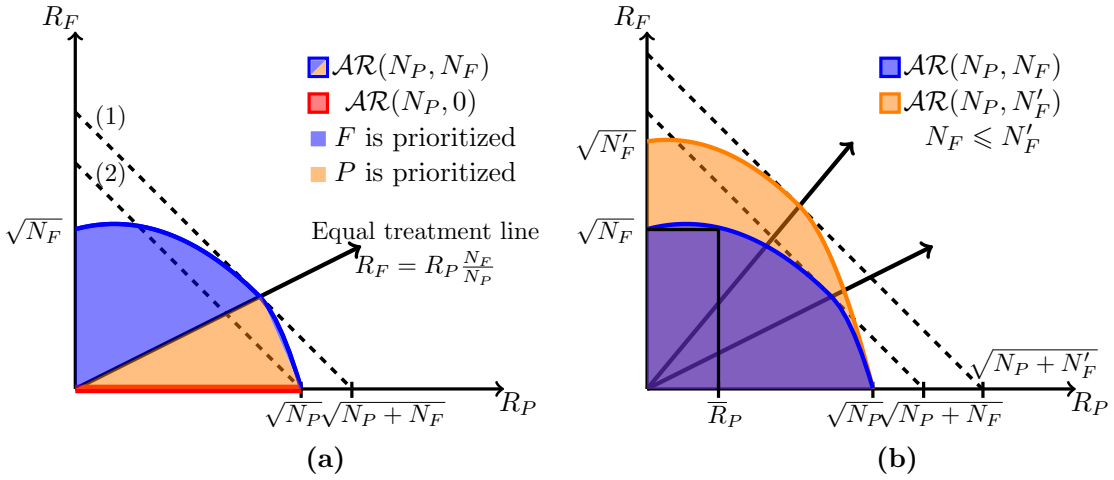


Figure C.1: Geometrical representation of achievable revenues set. We consider $\mathcal{AR}(N) = [0, \sqrt{N}]$ and a quadratic boundary for the sets $\mathcal{AR}(N_P, N_F)$.

We can use this representation to gain a better understating of Problem (4.3). In Figure C.1 (b), we can see that for different flexible supply hours (different values of N_F) there is a corresponding set of achievable revenues. Once we fix the flexible supply hours, it is possible to identify the optimal private supply revenue for a given flexible supply revenue. In the example of Figure C.1 (b), given flexible supply N_F (blue set) and for a fixed flexible supply revenue $R_F = \sqrt{N_F}$, we define $\overline{R}_P$ to be the largest possible achievable private supply revenue, as shown in the figure. This provides insights on how to solve Problem (4.3): for every pair (N_F, R_F) we can find an optimal private supply revenue by moving horizon-

tally until we hit the boundary of $\mathcal{AR}(N_P, N_F)$. However, we cannot do this for every pair (N_F, R_F). Consider again Figure C.1 (b), for the pair $(N_F, \sqrt{N_F})$ there exists at least one R_P (e.g., $\overline{R}_P$) such that $(R_P, R_F) \in \mathcal{AR}(N_P, N_F)$. But for the pair $(N_F, \sqrt{N'_F})$ there is no such R_P because $R_F = \sqrt{N'_F}$ is too large for the supply N_F and cannot be achieved with the the total supply hours available, $N_P + N_F$. We could, nevertheless, increase N_F to N'_F to be able to find a value of R_P such that (R_P, R_F) is achievable.

At an intuitive level, when solving Problem (4.3), the firm will always want the highest possible value of R_P given N_F, R_F and, therefore, will always choose a policy that guarantees $R_P = \overline{R}_P(N_F, R_F)$. Once this choice is made, the firm must optimize over the feasible pairs of (N_F, R_F) in $\mathcal{D}$. However, the firm must also take into account the equilibrium condition. In Figure C.2 (a), we depict the domain of the maximal private supply revenue function, $\mathcal{D}$, together with the equilibrium condition. For a given flexible supply revenue $\widetilde{R_F}$ the firm can choose a single value of flexible supply hours $\widetilde{N_F}$ which is consistent with the equilibrium condition. That is, the intersection of the domain, $\mathcal{D}$, with the equilibrium condition, $\gamma R_F = rN_F$, represent the space where the firm optimizes. In turn, we have transformed the firm's problem into a one-dimensional optimization problem in which we can optimize over the flexible supply revenue (the y-axis in Figure C.2 (a)). This discussion motivates Proposition 13.

Proposition 13 provides a simple characterization of the firm's problem as a one dimensional optimization problem. We can simply analyze changes in R_F which translate into movement along the equilibrium line. The latter also yields a change in N_F and, in turn, implies a change in the maximal private supply revenue function, $\overline{R}_P(N_F, R_F)$. This ultimately changes the firm's total profit, Profit$(\cdot)$. If there is profitable deviation around the optimal equal treatment policy that leads to a solution above or below the equal treatment line (see Figure C.2 (b)), then prioritization will naturally emerge.

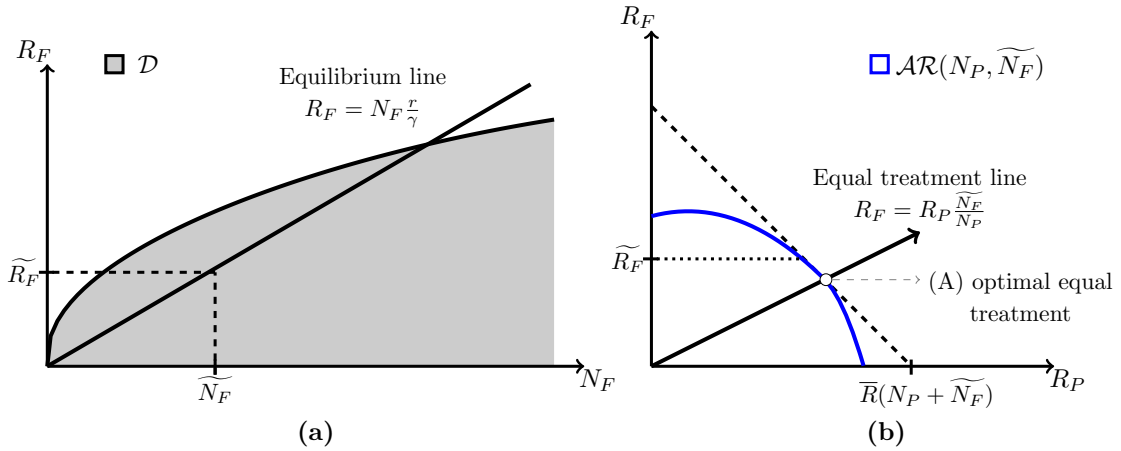


Figure C.2: **(a)** Domain of the maximal private supply revenue function and equilibrium line. **(b)** Achievable revenue set of $(\widetilde{N_F}, \widetilde{R_F})$.

C.6 Asymmetric supply

In Assumption 1, we assume supply agents are symmetric and have a similar performance in terms of the capability of acquiring revenues. For instance, a worker in a factory may have a similar work efficiency no matter whether it is a contractor or an employee. On the other hand, the symmetry assumption is also a simplification due to the fact of the complexity of comparing efficiencies. For the example of ride-hailing, riders may indeed be reluctant to use autonomous vehicles (AVs), and autonomous vehicles may be limited in some areas of a city so that the efficiency of AVs is negatively impacted on the demand side. However, on the supply side, it is also possible that the operation of AVs may be more efficient than human-driven vehicles (HVs) because of their automation. Martínez-Díaz and Soriguera (2018) point out AVs would lead to efficient traffic if a cooperative environment is built. Compared with human-driven vehicles, AVs may react faster and communicate more smoothly with the infrastructure, the cloud servers, and other vehicles. These different influences might cancel each other out, so the net effect on their efficiency is uncertain. Nevertheless, in this section, we discuss a method to extend our results to the case in which the supply agents

are asymmetric and have different efficiencies.

If the flexible supply and private supply are asymmetric, it means that they have different capabilities to achieve revenues, and the achievable revenues depend on the composition of the supply. With the same available hours, private supply may be able to earn higher or lower revenue than flexible supply. In other words, if private supply is more efficient than flexible supply, it means private supply can earn more revenue per unit of time. For example, ceteris paribus, an AV may earn more revenue per hour than an HV, because it is able to react faster and serves more customers per hour without needing any break. If we compare an AV with 10 HVs, however, the AV might be defeated as the revenue earned by these 10 HVs would be higher; and if we compare an AV with 5 HVs, the result might be a tie. In this sense, we can say the efficiency of an AV is equal to the efficiency of 5 HVs. This provides another way to build the relationship of the revenues between different mixes of supply. In Assumption 1, we assume the two types of supply are interchangeable and have the same efficiency so that a private agent is equivalent to a flexible agent. If the two types of supply are not interchangeable, we can consider that a private agent is equivalent to $\eta > 0$ number of flexible agents. In other words, the revenue that can be achieved by a private agent also can be achieved by η flexible agents. This motivates us to extend Assumption 1 as:

Assumption 4 (Asymmetry of supply types)**.** With a ratio $\eta > 0$, the feasible total revenues which can be achieved by a two-type policy can also be achieved by a single-type policy with private supply only, and vice versa:

$$\forall N_P, N_F \geqslant 0,\ \{R_P + R_F | (R_P, R_F) \in \mathcal{AR}(N_P, N_F)\} = \mathcal{AR}(\eta N_P + N_F)$$

Note that $\mathcal{AR}(N)$ now specifically refers to the set of revenues that is achievable by flexible supply only.

Here, η can be interpreted as an efficiency ratio between two types of supply agents. We suppose this η is independent of all other factors, such as the mix of supply and specific policies. In other words, we assume that each private agent is, ceteris paribus, able to acquire

the revenue that η flexible agents can make. In particular, if $\eta > 1$, private agents are more efficient and can obtain more per-hour revenue than the same number of flexible agents; if $\eta < 1$, private agents are less efficient and obtain less per-hour revenue than the same number of flexible agents. Notice that Assumption 1 is a special case of Assumption 4 with $\eta = 1$. We understand that Assumption 4 is still limited and cannot cover all the situations. And in practice, this efficiency ratio might also vary when the mix of supply is different. For example, the per-hour revenue earned by an independent AV may be equivalent to what earned by 5 HVs, but the per-hour revenue earned by an AV in a group of AVs may be equivalent to what earned by 10 HVs, as AVs can communicate with each other and cooperate without any internal conflict. Nonetheless, Assumption 4 greatly extends Assumption 1 to consider a possible asymmetry between the types by using a linear comparison for the revenues. The other forms of asymmetric effects such as a nonlinear relationship (e.g. η is a function of the mix.) are application specific and vary from case to case. And this limitation does not affect the insights which we illustrate behind the optimal prioritization and staffing strategy.

Now that Assumption 1 is extended to Assumption 4, we also need to modify the definition of equal treatment and prioritization. Due to the asymmetry, it is not appropriate to use average hourly revenue to define equal treatment and prioritization. Equal treatment should mean that the firm is agnostic to the types of supply and treats each agent by the exactly same policy. In the above paragraph, when we define η, we say that each private agent is, ceteris paribus, able to acquire the revenue that η flexible agents can make. This means that with the same policy, the average hourly revenue of each private agent should be equal to the average hourly revenue of η flexible agents. This motivates us to extend Definition 2 to the following Definition 6.

Definition 6 (Prioritization and Equal treatment for Asymmetric Supply)**.** With a ratio $\eta > 0$, for $N_P > 0$, we say that $(R_P, R_F) \in \mathcal{AR}(N_P, N_F)$

1. (Prioritizing Flexible Supply.) prioritizes flexible supply if and only if

$$N_F > 0 \text{ and } \frac{R_P}{\eta N_P} < \frac{R_F}{N_F},$$

2. (Prioritizing Private Supply.) and that it prioritizes private supply if and only if

$$N_F > 0 \text{ and } \frac{R_P}{\eta N_P} > \frac{R_F}{N_F}, \quad \text{or} \quad N_F = 0 \text{ and } \frac{R_P}{\eta N_P} > \frac{r}{\gamma}.$$

In any other case, we say that (R_P, R_F) satisfies equal treatment and we use $\mathcal{ET}(N_P, N_F) \subseteq \mathcal{AR}(N_P, N_F)$ to denote the set of equal treatment revenue pairs.

When we equally treat the two types of supply agents, the hourly revenues of private supply is proportional to flexible supply with the ratio η. For instance, a private agent may complete a request and become available faster than flexible agents, so they are able to obtain a higher revenue even if we ignore the types of supply. Notice that an equal treatment policy means the ignorance of types instead of fairness or equal revenue. Accordingly, Assumption 2 also needs to be adapted to Assumption 4 and Definition 6. That is, Assumption 2 becomes:

Assumption 5 (Equal-treatment policies can achieve any feasible revenue)**.** Given any supply $N_P \geqslant 0, N_F \geqslant 0$, any achievable revenue $R \in \mathcal{AR}(\eta N_P + N_F)$ is achievable by an equal treatment policy.

$$\exists\, (R_P, R_F) \in \mathcal{ET}(N_P, N_F),\ R_P + R_F = R.$$

Validation of all the results with the new assumptions and definition. To see all the results still hold, we can consider these modifications from another view. The new assumptions and definition is equivalent to defining $\widehat{N_P} = \eta N_P$ and $\widehat{C_P} = C_P/\eta$ and replacing the original N_P, C_P with $\widehat{N_P}, \widehat{C_P}$. Here, we can call $\widehat{N_P}$ as the equivalent available hour of private supply and $\widehat{C_P}$ as the equivalent operation cost of private supply, in the sense that the available hours of a private agent is computed as how long a flexible agent needs to complete the same amount of work which can be done by a private agent in N_P hours. For instance, in

an eight-hour working day system, if a private worker can produce 2 products per hour and a flexible worker can only produce 1 product per hour, the production of a private worker is equivalent to the production of a flexible worker who works for 16 hours per day. Also, $\widehat{C_P}$ is how much we need to pay for a private worker for the amount of work that can be done by a flexible worker per hour. For example, if the salary for a private worker (i.e. C_P) is \$ 10 dollars per hour, and the reserve earning of a flexible worker (i.e. r) is \$ 8 dollars per hour, the operation cost of private supply is actually lower because they have double productivity than flexible supply. Therefore, we can see that all the results will still be valid, if we replace N_P, C_P with $\widehat{N_P}, \widehat{C_P}$ everywhere.

Overall, the focus is to extend Assumption 1 and Definition 2 simultaneously and find a way to describe a relationship of the revenues between the two types of supply. As we explain above, we know these adjustments are still limited and the efficiency of supply might not be always comparable by a linear relationship. Nonetheless, we believe the insights we deliver are valid in most cases. A more complex asymmetric situation is another option for future study.

Bibliography

Afèche, P., Liu, Z., and Maglaras, C. (2023). Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. *Manufacturing & Service Operations Management*, 25(5):1890–1908.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction, Judgment, and Complexity: A Theory of Decision-Making and Artificial Intelligence*, pages 89–110. University of Chicago Press.

Alizamir, S., de Véricourt, F., and Wang, S. (2020). Warning against recurring risks: An information design approach. *Management Science*, 66(10):4612–4629.

Anderson, B. R., Shah, J. H., and Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. *Available at arXiv:2402.01536*.

Baidoo-Anu, D. and Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

Banerjee, S., Kanoria, Y., and Qian, P. (2018). State dependent control of closed queueing networks. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '18, page 2–4. Association for Computing Machinery.

Baron, O., Berman, O., and Nourinejad, M. (2022). Introducing autonomous vehicles: Adoption patterns and impacts on social welfare. *Manufacturing & Service Operations Management*, 24(1):352–369.

Bastani, H., Bastani, O., and Sinchaisri, W. P. (2022). Improving human decision-making with machine learning. *Available at arXiv:2108.08454*.

Benjaafar, S., Ding, J.-Y., Kong, G., and Taylor, T. (2022). Labor welfare in on-demand service platforms. *Manufacturing & Service Operations Management*, 24(1):110–124.

Benjaafar, S., Wang, Z., and Yang, X. (2023). The impact of automation on workers when workers are strategic: The case of ride-hailing. *Available at SSRN 3919411*.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer New York.

Bertsimas, D. (1995). The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing Systems*, 21:337–389.

Bertsimas, D. and Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Mathematics of Operations Research*, 21(2):257–306.

Besbes, O., Castro, F., and Lobel, I. (2021). Surge pricing and its spatial supply response. *Management Science*, 67(3):1350–1367.

Bhandari, A., Scheller-Wolf, A., and Harchol-Balter, M. (2008). An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. *Management Science*, 54(2):339–353.

Bhat, A., Agashe, S., Oberoi, P., Mohile, N., Jangir, R., and Joshi, A. (2023). Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 436–452.

Bimpikis, K., Candogan, O., and Saban, D. (2019). Spatial pricing in ride-sharing networks. *Operations Research*, 67(3):744–769.

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6).

Bommasani, R., Creel, K., Kumar, A., Jurafsky, D., and Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Boyacı, T., Canyakmaz, C., and de Véricourt, F. (2023). Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science*.

Brand, J., Israeli, A., and Ngwe, D. (2023). Using GPT for market research. *Available at SSRN 4395751*.

Braverman, A., Dai, J. G., Liu, X., and Ying, L. (2019). Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452.

Cachon, G. P., Daniels, K. M., and Lobel, R. (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384.

Cachon, G. P., Dizdarer, T., and Tsoukalas, G. (2021). Decentralized or centralized control of online service platforms: Who should set prices? *Available at SSRN 3957209*.

Cachon, G. P., Dizdarer, T., and Tsoukalas, G. (2022). Pricing control and regulation on online service platforms. *Available at SSRN 3957209*.

Castillo, J. C., Knoepfle, D., and Weyl, G. (2017). Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 241–242. Association for Computing Machinery.

Castro, F., Gao, J., and Martin, S. (2022). Supply prioritization in hybrid marketplaces. *SSRN 4119096*.

Chakravarty, A. K. (2021). Blending capacity on a rideshare platform: Independent and dedicated drivers. *Production and Operations Management*, 30(8):2522–2546.

Chaney, A. J. B., Stewart, B. M., and Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 224–232.

Chen, N., Hu, M., and Li, W. (2022). Algorithmic decision-making safeguarded by human knowledge. *Available at arXiv:2211.11028*.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51).

Dacre, M., Glazebrook, K., and Nino-Mora, J. (1999). The achievable region approach to the optimal control of stochastic systems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(4):747–791.

Dai, T. and Singh, S. (2023). Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. *Available at SSRN 3987454*.

Davalos, J. (2022). Uber revives self-driving taxi dreams, plans to start this year. Last accessed: 2024-03-05.

Dave, P. and Jin, H. (2021). Google self-driving spinoff waymo begins testing with public in san francisco. Last accessed: 2022-05-19.

de Véricourt, F. and Gurkan, H. (2023). Is your machine better than you? you may never know. *Management Science*.

de Véricourt, F., Gurkan, H., and Wang, S. (2021). Informing the public about a pandemic. *Management Science*, 67(10):6350–6357.

Denny, P., Kumar, V., and Giacaman, N. (2023). Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2023, page 1136–1142.

DiDi (2023). Didi autonomous driving plans to introduce its first mass-produced robotaxi to didi's ride-hailing platform by 2025. Last accessed: 2024-03-05.

Dolan, S. (2022). Crowdsourced delivery explained: making same day shipping cheaper through local couriers. Last accessed: 2022-05-19.

Dong, J. and Ibrahim, R. (2020). Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research*, 68(4):1238–1264.

Doshi, A. R. and Hauser, O. (2024). Generative artificial intelligence enhances creativity but reduces the diversity of novel content. *Available at SSRN 4535536*.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *Available at arXiv:2303.10130*.

Fagnant, D. J. and Kockelman, K. M. (2018). Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. *Transportation*, 45(1):143–158.

Fatehi, S. (2024). The path to green in ride-hailing. *Available at SSRN 4722757*.

Fernandes, P. and Nunes, U. (2010). Platooning of autonomous vehicles with intervehicle communications in sumo traffic simulator. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 1313–1318.

Freund, D., Lobel, I., and Zhao, J. (2022). On the supply of autonomous technologies in open platforms. *Available at SSRN 4178508*.

Fuller, J. B., Raman, M., Palano, J., Bailey, A., Vaduganathan, N., Kaufman, E., Laverdière, R., and Lovett, S. (2020). Building the on-demand workforce. Last accessed: 2022-05-19.

Gallager, R. G. et al. (2008). *Principles of digital communication*, volume 1. Cambridge University Press Cambridge, UK.

Gentzkow, M. and Kamenica, E. (2014). Costly persuasion. *American Economic Review*, 104(5):457–62.

Github (2023). Github copilot · your AI pair programmer. Last accessed: 2024-02-05.

Gurvich, I., Lariviere, M., and Moreno, A. (2019). *Operations in the On-Demand Economy: Staffing Services with Self-Scheduling Capacity*, pages 249–278. Springer International Publishing, Cham.

Hall, J., Horton, J., and Knoepfle, D. T. (2021). Pricing in designed markets: The case of ride-sharing. *Working paper*.

Hartmann, J., Schwenzow, J., and Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *Available at arXiv:2301.01768*.

Haviv, M. (2013). *Queues–A Course in Queueing Theory*. Springer, New York, NY.

Hawkins, A. J. (2021). Amazon's zoox will test its autonomous vehicles on seattle's rainy streets. Last accessed: 2022-05-19.

Hazan, J., Lang, N., Chua, J., Doubara, X., Steffens, T., and Ulrich, P. (2016). Will autonomous vehicles derail trains? Last accessed: 2024-03-05.

He, E. J. and Goh, J. (2022). Profit or growth? dynamic order allocation in a hybrid workforce. *Management Science*, 68(8):5891–5906.

Hu, B., Hu, M., and Zhu, H. (2022). Surge pricing and two-sided temporal responses in ride hailing. *Manufacturing & Service Operations Management*, 24(1):91–109.

Hu, K. (2023). ChatGPT sets record for fastest-growing user base. Last accessed: 2024-02-05.

Hu, M., Wang, J., and Zhang, Z. (2023). Implications of worker classification in on-demand economy. *Available at SSRN 4076484*.

Hu, M. and Zhou, Y. (2020). Price, wage, and fixed commission in on-demand matching. *Available at SSRN 2949513*.

Hu, M. and Zhou, Y. (2022). Dynamic type matching. *Manufacturing & Service Operations Management*, 24(1):125–142.

Hur, K. (2022). Lyft plans to build a hybrid network of autonomous and driver vehicles, co-founder says. Last accessed: 2024-03-05.

Hussain, R. and Zeadally, S. (2019). Autonomous cars: Research results, issues, and future challenges. *IEEE Communications Surveys Tutorials*, 21(2):1275–1313.

Ibrahim, R. (2018). Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management*, 27(2):234–250.

Ibrahim, R., Kim, S.-H., and Tong, J. (2021). Eliciting human judgment for prediction algorithms. *Management Science*, 67(4):2314–2325.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15.

Jiang, J. (2019). More americans are using ride-hailing apps. Last accessed: 2024-03-05.

Kalliamvakou, E. (2023). Research: quantifying github copilot's impact on developer productivity and happiness. Last accessed: 2024-02-05.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kanoria, Y. and Qian, P. (2020). Blind dynamic resource allocation in closed networks via mirror backpressure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20. Association for Computing Machinery.

Kesavan, S., Staats, B. R., and Gilland, W. (2014). Volume flexibility in services: The costs and benefits of flexible labor resources. *Management Science*, 60(8):1884–1906.

Khosrowshahi, D. (2020). The high cost of making drivers employees. Last accessed: 2022-05-19.

Kinsella, B. (2023). OpenAI to offer ChatGPT customization and shares bias guidelines. Last accessed: 2024-02-05.

Krishnan, V., Iglesias, R., Martin, S., Wang, S., Pattabhiraman, V., and Van Ryzin, G. (2022). Solving the ride-sharing productivity paradox: Priority dispatch and optimal priority sets. *INFORMS Journal on Applied Analytics*, 52(5):433–445.

Lanzetti, N., Schiffer, M., Ostrovsky, M., and Pavone, M. (2023). On the interplay between self-driving cars and public transportation. *IEEE Transactions on Control of Network Systems*.

Lian, Z. and van Ryzin, G. (2022). Capturing the benefits of autonomous vehicles in ride-hailing: The role of dispatch platforms and market structure. *Available at SSRN 3716491*.

Lingard, L. (2023). Writing with chatgpt: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education*, 12(1):261–270.

Litman, T. (2023). Autonomous vehicle implementation predictions: Implications for transport planning. Technical report, Victoria Transport Policy Institute.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Lobel, I., Martin, S., and Song, H. (2024). Frontiers in operations: Employees vs. contractors: An operational perspective. *Manufacturing & Service Operations Management*, 26(4):1306–1322.

Lu, L., Fang, X., Feng, G., and Savin, S. (2024). Taxis on ride-hailing platforms: Managing on-demand urban mobility ecosystems. *Available at SSRN 4771777*.

Ludlow, E. (2023). Amazon's driverless robotaxis take to las vegas streets. Last accessed: 2024-03-05.

Lyft (2021). The key to av deployment: the rideshare network. Last accessed: 2022-11-01.

Lyft (2024). The future of transportation is in your pocket. Last accessed: 2024-03-05.

Ma, Y., Wang, Z., Yang, H., and Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA Journal of Automatica Sinica*, 7(2):315–329.

Martínez-Díaz, M. and Soriguera, F. (2018). Autonomous vehicles: theoretical and practical challenges. *Transportation Research Procedia*, 33:275–282. XIII Conference on Transport Engineering, CIT2018.

Mclaughlin, B. and Spiess, J. (2023). Algorithmic assistance with recommendation-dependent preferences. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, page 991. Association for Computing Machinery.

Midjourney (2023). Midjourney. Last accessed: 2024-02-05.

Mirzaeian, N., Cho, S.-H., and Scheller-Wolf, A. (2021). A queueing model and analysis for autonomous vehicles on highways. *Management Science*, 67(5):2904–2923.

Mok, A. (2023). 'Prompt engineering' is one of the hottest jobs in generative AI. here's how it works. Last accessed: 2024-02-05.

Motoki, F., Neto, V. P., and Rodrigues, V. (2023). More human than human: measuring chatgpt political bias. *Public Choice*.

Narayanan, S., Chaniotakis, E., and Antoniou, C. (2020). Shared autonomous vehicle services: A comprehensive review. *Transportation Research Part C: Emerging Technologies*, 111:255–293.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Ogg, M. (2021). "world's largest online workforce" freelancer.com sees enterprises in desperate need of talent. Last accessed: 2022-05-19.

OpenAI (2023a). Custom instructions for ChatGPT. Last accessed: 2024-02-05.

OpenAI (2023b). Introducing ChatGPT. Last accessed: 2024-02-05.

OpenAI (2023c). ChatGPT can now see, hear, and speak. Last accessed: 2024-02-05.

Ostrovsky, M. and Schwarz, M. (2019). Carpooling and the economics of self-driving cars. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19. Association for Computing Machinery.

Padmakumar, V. and He, H. (2024). Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*.

Pugh, C. C. (2015). *Real Mathematical Analysis*. Springer Cham.

Reed, S., Campbell, A. M., and Thomas, B. W. (2022). The value of autonomous vehicles for last-mile deliveries in urban environments. *Management Science*, 68(1):280–299.

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, 12(3).

Saatci, Y. and Wilson, A. G. (2017). Bayesian gan. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sallam, M. (2023). Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6).

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *Available at arXiv:2305.17493*.

Siddiq, A. and Taylor, T. A. (2022). Ride-hailing platforms: Competition and autonomous vehicles. *Manufacturing & Service Operations Management*, 24(3):1511–1528.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Solis, S. (2021). Should uber, lyft drivers be employees? tech companies, labor rights advocates square off over how ride-sharing drivers are classified. Last accessed: 2022-05-19.

Taylor, T. A. (2018). On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4):704–720.

Thompson, S. A. (2023). Uncensored chatbots provoke a fracas over free speech. Last accessed: 2024-02-05.

Uber (2016). Pittsburgh, your self-driving uber is arriving now. Last accessed: 2022-05-19.

Uber (2024). Shaping the future of transportation and delivery. Last accessed: 2024-03-05.

Vaithilingam, P., Zhang, T., and Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22. Association for Computing Machinery.

Wang, G., Zhang, H., and Zhang, J. (2022). On-demand ride-matching in a spatial model with abandonment and cancellation. *Operations Research*.

Waymo (2024). Waymo one. Last accessed: 2024-03-05.

Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Wei, C., Xie, S. M., and Ma, T. (2021). Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Whitt, W. (1984). Open and closed models for networks of queues. *AT&T Bell Laboratories Technical Journal*, 63(9):1911–1979.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2022). An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2020). Fine-tuning language models from human preferences. *Available at arXiv:1909.08593*.

Özkan, E. and Ward, A. R. (2020). Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70.