Essays on the Economics of Algorithms, Markets, and Organizations

by

Lindsey Raymond

B.A. Economics, Yale University, 2012
S.M. Management Research, Massachusetts Institute of Technology, 2019

Submitted to the Department of Management
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Lindsey Raymond Department of Management May 3, 2024
Certified by:	Danielle Li Associate Professor of MIT Sloan School of Management, Thesis Supervisor
Certified by:	Sendhil Mullainathan University of Chicago, Roman Family University Professor, Thesis Supervisor
Accepted by:	Eric So Sloan Distinguished Professor of Global Economics and Management Faculty Chair, MIT Sloan PhD Program

Essays on the Economics of Algorithms, Markets, and Organizations

by

Lindsey Raymond

Submitted to the Department of Management
on May 3, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

This dissertation contains three chapters that study how digitization and increasingly reliance on algorithms shapes workers, organizations, and markets. In the first chapter, I show how the digitization of public housing records leads to the entry of investors using algorithms. Digitization and entry lead to changes in equilibrium prices and allocation in the US residential real estate market. Consistent with a theoretical model of comparative advantage, I observe shifts in investment patterns for both humans and algorithmic investors and changing house prices, particularly for minority homeowners. In the second chapter, I study how hiring algorithm design shapes the effects of algorithms in the labor market. Using data from a professional services firm, I show that incorporating exploration can improve the quality of the interview screening process (as measured by eventual hiring rates), while also increasing demographic diversity, relative to the firm's existing practices. While the adoption of automated approaches to hiring is often associated with decreasing access to opportunity, we show the impact on efficiency and equity depends on algorithm design choices. In the third chapter, joint with Danielle Li and Erik Brynjolfsson, we study the staggered introduction of a generative AI-based conversational assistant using data from 5,000 customer support agents. Access to the tool increases productivity, as measured by issues resolved per hour, by 14% on average, including a 34% improvement for novice and low-skilled workers but with minimal impact on experienced and highly skilled workers. We provide suggestive evidence that the AI model disseminates the best practices of more able workers, helps newer workers move down the experience curve, and improves worker learning. Our results suggest that access to generative AI can increase productivity, with large heterogeneity in effects across workers. Together, these chapters highlight how the increasing prevalence of algorithmic decision making impacts workers, firms, and markets.

Thesis supervisor: Danielle Li
Title: Associate Professor of MIT Sloan School of Management

Thesis supervisor: Sendhil Mullainathan
Title: University of Chicago, Roman Family University Professor

Acknowledgments

I must extend my most heartfelt appreciation to my family and partner for their relentless belief, constant encouragement, and unwavering support during graduate school. I am deeply grateful to my committee chairs, Danielle Li and Sendhil Mullainathan, for their inordinately patient mentorship and thoughtful guidance, shaping me into the scholar and person I am today. My sincere thanks also go to my committee members, Erik and Scott, for their support and encouragement, as well as many other faculty members, both at MIT and elsewhere, who were instrumental throughout my journey. I am also deeply indebted to the friends I have made during the PhD, both at MIT and elsewhere, who were infectiously enthusiastic, generous, and kind, and inspired me to persist.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Decision making is a central economic activity. Firms decide how to allocate resources, who to hire, and whether to enter or exit markets. Workers must adapt to unexpected events, set priorities, and delegate tasks. Traditionally, decision making has been the province of human expertise: people make decisions based on anything from gut instinct to systematic processes within organizations. However, due to the growing availability of digitized information and advances in machine learning, algorithms (automated procedures making inferences from data) are increasingly central to decision making.

This dissertation studies how the growing digitization of economic activity and the increasing reliance on algorithms impacts workers, firms, and markets. The first chapter studies how the digitization of public housing records changes the US residential housing market. In the second chapter, I illustrate how the design of algorithms plays a pivotal role in determining how automation impacts hiring decisions. In the third chapter, I study how the introduction of a generative AI-based assistance impacts worker productivity and the experience of work.

1.0.1 The Market Effects of Algorithms

In the first chapter, I explore the equilibrium effects of digitization and algorithm adoption on the US housing market. While there is excitement about the potential of algorithms to optimize individual decision-making, changes in individual behavior will, almost inevitably, impact markets. Yet little is known about these effects. In this paper, I study how the availability of algorithmic prediction changes entry, allocation, and prices in the US residential real estate market, a key driver of household wealth. I identify a *market-level* natural experiment that generates variation in the cost of using algorithms to value houses: digitization, the transition from physical to digital housing records. I show that digitization leads to entry by investors using algorithms, but does not push out investors using human judgment. Instead, human investors shift towards houses that are difficult to predict algorithmically. Algorithmic investors predominantly purchase minority-owned homes, a segment of the market where humans may be biased. Digitization increases the average sale price of minority-owned homes by 5% or $5,000 and nearly eliminates racial disparities in home prices. Algorithmic investors, via competition, affect the prices paid by humans for minority homes, which drives most of the reduction in racial disparities. This decrease in racial inequality underscores the

potential of algorithms to mitigate human biases at the market level.

1.0.2 Hiring as Exploration

In joint work with coauthors Danielle Li and Peter Bergman, the second chapter views hiring as a contextual bandit problem: to find the best workers over time, firms must balance "exploitation" (selecting from groups with proven track records) with "exploration" (selecting from under-represented groups to learn about quality). Yet modern hiring algorithms, based on "supervised learning" approaches, are designed solely for exploitation. Instead, we build a resume screening algorithm that values exploration by evaluating candidates according to their statistical upside potential. Using data from professional services recruiting within a Fortune 500 firm, we show that this approach improves the quality (as measured by eventual hiring rates) of candidates selected for an interview, while also increasing demographic diversity, relative to the firm's existing practices. The same is not true for traditional supervised learning based algorithms, which improve hiring rates but select far fewer Black and Hispanic applicants. In an extension, we show that exploration-based algorithms are also able to learn more effectively about simulated changes in applicant hiring potential over time. Together, our results highlight the importance of incorporating exploration in developing decision-making algorithms that are potentially both more efficient and equitable.

1.0.3 Generative AI at Work

In the third chapter - with coauthors With coauthors Danielle Li and Erik Brynjolfsson - I explore the impact of new generative AI tools on worker productivity. New AI tools have the potential to change the way workers perform and learn, but little is known about their impacts on the job. In this paper, we study the staggered introduction of a generative AI-based conversational assistant using data from 5,179 customer support agents. Access to the tool increases productivity, as measured by issues resolved per hour, by 14% on average, including a 34% improvement for novice and low-skilled workers but with minimal impact on experienced and highly skilled workers. We provide suggestive evidence that the AI model disseminates the best practices of more able workers and helps newer workers move down the experience curve. In addition, we find that AI assistance improves customer sentiment, increases employee retention, and may lead to worker learning. Our results suggest that access to generative AI can increase productivity, with large heterogeneity in effects across workers.

1.0.4 Conclusion and Future Directions

These three essays illustrate that the digitization of information and the growing use of algorithms are changing markets, impacting both workers and firms. More work is needed not only to better understand the effects of these tools, but also to integrate our understanding of economic behavior with computation in order to build tools that improve welfare.

Chapter 2

The Market Effects of Algorithms

Many consequential decisions depend on predictions — hiring depends on predictions of who will be most productive; extending credit depends on the prediction of default; and investing decisions crucially rely on predictions of returns. The advent of machine learning and digital data has created a great deal of interest in the use of technology in prediction problems. Prediction technologies, or algorithms, could potentially change and improve decision making. In fact, a rich literature has already explored a variety of questions, including whether algorithms do better than humans or when algorithms might inherit biases of people making such decisions. Although much progress has been made, one area remains largely unexplored: *market* level impacts.

In addition to enhancing individual decision making, algorithms could have broader market-level and equilibrium consequences. Algorithms could lead to new entrants, change the nature of competition, and alter market-level outcomes such as prices. These broader market dynamics mean that even if algorithms help people make better decisions, individuals could still be worse off. In contrast, even if algorithms do not uniformly improve individual decisions, sectors of the market where human error is most pronounced could benefit. Studies designed to capture the impact of algorithms on decision quality cannot, by design, account for effects beyond the individual (or firm) level.

I empirically examine the market effects of algorithms in the US residential real estate market. In particular, I focus on investors, who buy houses to rent out or resell. I study the housing market due to the importance of the setting, the centrality of prediction in decision making, and the presence of an elegant natural experiment. First, housing is the largest contributor to the wealth of the median household and the largest asset market and therefore is a substantively interesting application in its own right (Derenoncourt et al., 2022; Malone, 2023).[1] Second, investing in real estate hinges on accurate predictions: investors aim to forecast potential rental income, property appreciation, annual maintenance costs, and overall future value. Yet, prediction is a difficult cognitive task for humans. Various human behavioral biases, such as the influence of weather, sentiment, anchoring, and loss aversion, among other heuristics, have been shown to impact housing prices (Busse et al., 2012; Kermani and Wong, 2021*a*; Salzman and Zwinkels, 2017).[2] Human investors also spend a lot of effort trying to mitigate the influence of these behavioral biases.

To address the central identification challenge—that algorithms are not randomly assigned to markets—I exploit an institutional feature of real estate markets and a simple yet fundamental insight: machine learning algorithms require machine-readable data. Specifically, algorithms need detailed property information, such as the number of bedrooms and bathrooms, yard size, age of the house, historical sale prices, home improvements records, and current market data. In the US, county governments are responsible for collecting this public information for routine administrative tasks like planning, legal processes, and taxation. This information was traditionally stored as paper documents and microfilm records in

[1] A single-family home is a structure designed for a single one household, usually on its own plot of land, and while often associated with homeownership, is the largest single segment of the rental market (Freddie Mac Economic & Housing Research, 2018; Neal, Goodman and Young, 2020).

[2] Throughout the paper, I use "algorithms", "machine-generated predictions" and "ML algorithms" interchangeably. I also use "algorithmic investor", "investor using algorithms" and "investor using algorithmic valuation" to refer to investors using algorithms and "human investors" to refer to those who rely on human expertise.

county offices. However, as part of a broader move towards open and transparent governance, counties began to digitize these archives into electronic database systems. This transition from physical to digital records—a process known as digitization—generates variation in the cost of using algorithms to value property. I analyze the county-level transition to digitized housing data across Georgia, North Carolina, South Carolina and Tennessee, over the period from 2009 to 2021. This natural experiment allows me to contribute some of the first evidence on the market-level effects of algorithms.[3]

First, I investigate whether digitization actually prompts entry by investors making use of predictive algorithms—called algorithmic investors.[4] After digitization, there is a six-fold increase in the number of houses bought by algorithmic investors. This surge is sharp and evident in the raw data and in the event study analysis. On average, algorithmic investors account for roughly 10% of all investor activity after digitization.

Digitization also changes the structure of the investor market. Before digitization, the market was dominated by small "mom-and-pop" entrepreneurs. The average human investor would typically purchase around 1.5 houses annually, operating within two different zip codes. Many of these investors worked as contractors, plumbers, or real estate agents, often buying in their local area. In general, this market was believed to remain localized and fragmented due to the considerable advantage of the mom-and-pop entrepreneurs in local and qualitative information and relevant expertise (Fields, 2018). Algorithmic investors, who depend on automated valuations instead of human expertise, are capable of buying hundreds of homes annually in hundreds of zip codes. Consequently, digitization has resulted in a 23% increase in the size of the average firm and has doubled the number of geographic areas in which the average firm invests.

These initial findings could be misleading if the timing of county digitization is correlated with the activities of algorithmic investors or overall housing market activity. Specifically, there are two principal potential confounders to consider. The first concern is that algorithmic investment firms themselves might influence when a county chooses to digitize its records. For example, if an investor wanted to buy houses within a certain county, they might lobby for or financially support improvements in record management. To address this concern, I check for any suggestion of preexisting investor interest and the timing of digitization. If the timing of digitization were indeed driven by investors, I would expect early-digitizing counties to differ from those digitizing later. However, the timing of digitization is not strongly correlated with any observable characteristics of counties.

Second, an unobserved factor, such as changing county business policies, could simultaneously affect both digitization and housing market activities. For instance, the construction of a new manufacturing plant could prompt a push for modernization within county administrations—leading to record digitization—and draw interest from algorithmic investors. To test for potential confounding by unobserved variables, I leverage bureaucratic inconsistencies in the timing of digitization for each property. Due to budgetary limitations,

[3]Prior work in this field often concentrates on issues such as the implications of algorithmic pricing for competitive dynamics or analyzes the broad impact of algorithmic trading in financial markets Aggarwal and Thomas (2014); Brown and MacKay (2023); Calder-Wang and Kim (2023); Calvino and Fontanelli (2023); Clark et al. (2023). I am unaware of other work on the market effects of ML-powered algorithms.

[4]I define algorithmic investor as an investor using algorithms to value houses. Investors not using algorithms are "human investors."

counties often digitized their records in batches, resulting in variability when each property became digitally accessible. With houses still awaiting digitization subject to the same potential unobserved variables—like changes in business policies—yet not being algorithmically assessable, these houses provide a natural control group.

To test for any evidence of confounders such as changing economic policies, I conduct a series of falsification tests at the county, house, and neighborhood levels and a triple-difference analysis. In the triple-difference analysis, I compare not-yet-digitized houses to digitized houses, before and after county digitization, in each county. In the falsification tests, I compare the impact of county digitization on digitized and not-yet-digitized houses within the same census tract, block group and block. Across all specifications, the observed effects are concentrated on properties that have been digitized, rather than those yet to be digitized. These patterns do not suggest substantial unobserved county- or neighborhood-level shocks driving the results. Overall, this suggests that algorithmic investors' activity depends on access to machine-readable data.

My second result investigates the natural question around digitization and the subsequent entry of algorithmic investors into the market: do these entities displace the small-scale, individual entrepreneurs who previously dominated the market? In fact, both traditional "mom-and-pop" entrepreneurs and algorithmic investors coexist in the post-digitization period. To make sense of this, I propose a conceptual framework that recognizes that humans and algorithms possess *distinct* comparative advantages. Humans have access to a wealth of non-digitizable information that is inaccessible to an algorithm. Humans see details such as the aesthetics of bathroom tiles, yard sunlight exposure, and ambient neighborhood noise levels. While limited to quantifiable data, algorithms derive their predictions from structured statistical relationships (Kahneman, Sibony and Sunstein, 2021; Mullainathan and Obermeyer, 2021). Furthermore, algorithms may not be susceptible to some sources of human errors, such as cognitive limitations or explicit prejudices. If the private information available to humans is important, then humans may do better. If humans make systematic mistakes, algorithms may have an advantage. This framework generates two clear predictions: Human investors should specialize where private information is important and where their comparative advantage is strongest. Algorithms should target properties where human errors are most prevalent. I empirically explore each of these hypotheses in turn.

In line with my theoretical framework, digitization leads human investors to shift their focus toward properties that pose difficulties for algorithmic prediction and away from those that are easily predictable. To capture where models do well and where they struggle, I use commonly available house data to predict price with an extreme gradient boosted tree model. Houses vary widely in their predictability. The discrepancy between actual and algorithm-predicted prices is as high as 50% for some houses. For others, the model error is less than 1%. Using this measure, digitization doubles the likelihood that human investors purchase the hardest-to-predict properties and reduces by half their propensity to purchase the most predictable houses. These results are consistent with human investors gravitating toward parts of the market where their informational advantage is highest.

What makes some houses easy to predict with a model and others so difficult? For some properties, important information is missing from the digitized data sets, leading to significant model discrepancies. In others, the unobserved information matters more. Take older homes as an example: they often have hidden issues like lead paint, presenting a serious

health risk and entailing additional remediation costs. Since information about the presence of lead paint isn't accessible to algorithms—yet can be inferred or discovered by humans through inspection—models typically have a more difficult time valuing older properties. More recent houses built after the prohibition of materials like lead-based paints present fewer such challenges for algorithmic assessments. As a third example, data inaccuracies, such as errors in recording the number of bedrooms, disrupt accurate valuations with algorithms. While data errors cause problems for algorithms, human investors, who physically inspect properties and can count the number of bedrooms, will not be affected. These insights suggest a testable implication of my framework: Algorithmic investors should steer clear of properties where certain institutional factors enhance the informational advantage of humans. Using data errors, building regulations, and county zoning rules as illustrative examples, I show that algorithmic investors avoid houses with these characteristics, while humans invest in these houses.

In this framework, human errors create opportunities for algorithms. Although a variety of human behavioral biases could possibly generate human error, I will focus on racial bias (Whittle et al., 2014). Prior to the Fair Housing Act, race was explicitly used to determine house values. While it is now illegal to explicitly incorporate race, racial disparities remain and have been the subject of extensive study (Cutler, Glaeser and Vigdor, 1999; Elster and Zussman, 2022; Freddie Mac Economic & Housing Research, 2021; Perry, Rothwell and Harshbarger, 2018). Specifically, prior work has identified a persistent valuation gap: minority homeowners tend to receive lower prices for their homes compared to White homeowners, even after adjusting for house and neighborhood characteristics (Elster and Zussman, 2022; Harris, 1999; Perry, Rothwell and Harshbarger, 2018). This *race penalty* could be evidence that humans are undervaluing minority homes or driven by omitted variables or preferences.

Before exploring where algorithms buy houses, I estimate the race penalty in my sample prior to digitization and test for alternative explanations. Before digitization, a minority homeowner receives about 5% less than a White homeowner when selling their home, adjusting for house and neighborhood characteristics. While human biases could account for such disparities, this gap could also be driven by omitted variables that are correlated with homeowner race. A leading concern is that minority homeowners are often more cash-constrained or less wealthy, leading to differences in home maintenance or yard care reflected in the appearance of the house (Harris, 1999; Perry, Rothwell and Harshbarger, 2018). To investigate whether this gap simply reflects omitted variables associated with home appearance, I utilize a deep learning model trained on images of house exteriors, yards, and driveways. After controlling for aspects of house quality captured by house images, the race penalty persists. This suggests that differences in property maintenance or aesthetic factors do not fully explain the lower price received by minority homeowners.

Consistent with the possibility that humans may undervalue minority-owned homes, algorithmic investors disproportionately buy minority-owned homes. The impact of digitization on the likelihood of purchase by an algorithmic investor is six times larger for a minority homeowner compared to a White homeowner in the same census block. In other words, digitization of minority homes leads to a 250% increase in the probability that an algorithmic investor buys that home compared to 40% for a White-owned house. Moreover, areas where algorithmic investors are active do not have much larger shares of minority residents than those where human investors buy houses. This suggests that algorithmic investors tar-

get minority-owned houses, rather than just neighborhoods with higher shares of minority residents.

In my third set of results, I investigate how these changes in market composition impact overall prices and racial disparities in prices. Before digitization, both owner-occupiers and human investors typically pay around 5% less for homes owned by minorities compared to those owned by white homeowners. However, algorithmic investors, who enter and buy houses after digitization, do not exhibit a race penalty. In other words, the price algorithmic investors pay for a house does not depend on the race of the homeowner. After digitization, the race-associated price discrepancy decreases significantly and eventually disappears within six years of a county transitioning to digitized records. That is to say, before digitization, observably similar homes sell for different prices based on the seller's race. After digitization, observably similar houses sell for the same price.

Importantly, human investors and owner-occupiers drive much of this reduction in market-level racial disparities. After digitization, the race penalty among owner-occupier purchases decreases from 5% to 3%. Among human investors, the race penalty falls from 5% to 1.5%. This decline can be attributed to two main factors. First, algorithmic investors' presence may drive up house prices through competitive bidding, affecting final sale prices even in transactions they do not win. Second, transaction prices inform listing prices for new homes on the market; higher starting prices lead to higher sale prices for minority-owned homes, regardless of algorithmic investor participation. Given that owner-occupiers represent about 80% of the market, these indirect effects drive the overall reduction in racial pricing disparities. As a result of these changes, the aggregate impact of digitization is a 5% increase in average sale prices for homes owned by minorities, compared to a 1% increase for White-owned homes. These findings highlight how market interactions can amplify the impacts of algorithms in ways that firm-level analysis cannot capture.

Although one explanation for the increasing prices of minority homes might be human mistakes, another possibility is that algorithmic investors are simply overpaying for minority-owned houses. Algorithms do not see all aspects of house quality available to humans—potentially leading to adverse selection. To investigate whether this is occurring, I examine the gross returns—the difference between purchase and resale prices. If algorithmic investors consistently overpaid for minority-owned properties, their gross returns should be lower than the gross returns on White-owned homes. In fact, the gross returns on White and minority-owned homes are not statistically different. This holds true whether I use the resale price or an alternative measure of value, such as tax assessor estimates, to calculate gross margins.

Further undermining the adverse selection hypothesis is the increasing price that human buyers pay for minority-owned homes after digitization. If minority-owned homes were unobservably bad, human buyers should not be willing to pay more for these properties. Instead, the price human buyers are willing to pay for minority homes also rises after digitization. Together, this evidence is more consistent with humans previously undervaluing minority-owned than with algorithms overvaluing such properties.

So far, my analysis has focused on the transaction prices of the sold properties. Nonetheless, it's crucial to consider how these shifts might influence the valuation of unsold homes—assets that constitute a substantial proportion of wealth for the median household. My estimates suggest that digitization leads to a 6% appreciation in the average value of unsold minority-owned homes. This appreciation is considerable when viewed in relation to median

households' wealth: a 6% rise in property values corresponds to an increase by roughly 20% and 13% of the median Black and Hispanic family wealth, respectively (Bhutta et al., 2020).

These findings highlight how markets can amplify the effects of algorithms. Here, low average levels of algorithmic investor activity induce significant changes that impact those using algorithms and those not using algorithms alike. Competition and price effects lead to a reduction in racial disparities in property values, affecting homeowners across the market, and change the behavior of human investors and owner-occupiers. My findings are similar in spirit to Becker (1957), where competition penalizes and drives out firms with discriminatory views. The magnitude and patterns of these effects raise questions about how algorithms could be reshaping other parts of the economy.

This paper builds on a growing empirical literature on the impacts of access to algorithmic recommendations. Comparing human decision-makers' choices with predictive models has a long history (Dawes, 1971; Dawes, Faust and Meehl, 1989; Hastie and Dawes, 2001*a*). Modern advances in ML, increased computing power, and data availability have renewed interest in these questions. I build on prior work that shows that algorithmic recommendations can lead to, for example, improvements ranging from better heart attack diagnosis, to more efficient bail and hiring decisions.[5] Other work shows that access to algorithms translates into productivity or efficiency.[6] However, not all studies find positive effects.[7]

There is limited work on the effects of ML-powered algorithms at the market level. Brown and MacKay (2023); Calder-Wang and Kim (2023); Calvano et al. (2020); Clark et al. (2023) focus on the impact of ML-powered pricing algorithms on collusive behavior and price levels. Other studies focus on the impacts of automated algorithmic trading on the liquidity and pricing efficiency of financial markets (Chaboud et al., 2014; Hendershott, Jones and Menkveld, 2011; Upson and Van Ness, 2017). I examine the market-level impacts of algorithmic prediction outside of financial market trading.

This paper is closely related to a large literature on racial disparities in the housing market. Although centralized discrimination has declined over time, audit studies, surveys, and empirical work continue to find evidence consistent with racial discrimination in the housing market.[8] Racial disparities in house values contribute to the large and persistent

[5]For example, see Autor and Scarborough (2008); Frankel (2021); Li, Raymond and Bergman (2020); OECD (2023); Raghavan et al. (2020); The White House (2022) for applications in the labor market, Arnold, Dobbie and Yang (2018); Blattner and Nelson (2021); Einav, Jenkins and Levin (2013); Fuster et al. (2022); Gillis and Spiess (2019) for consumer finance, Abaluck et al. (2020); Chouldechova et al. (2018); Kleinberg et al. (2017*a*); Kleinberg, Mullainathan and Raghavan (2016); Mullainathan and Rambachan (2023); Mullainathan and Obermeyer (2021); Obermeyer and Emanuel (2016) for examples in the criminal justice system, health care, among other areas. See Kleinberg et al. (2017*b*, 2015); Rambachan (2022) for issues comparing human and machine predictions.

[6]See Brynjolfsson, Raymond and Li (2023) for the impacts of generative AI on productivity in customer service, Harris and Yellen (2023) for the impact of the adoption of predictive maintenance on repair costs in a trucking company. See Bubeck et al. (2023); Choi and Schwarcz (2023); Noy and Zhang (2023); Peng et al. (2023*a*) for additional effects of AI access on productivity, writing, and test taking capabilities.

[7]For instance, Acemoglu et al. (2022) finds no detectable relationship between AI investments and firm performance, while Babina et al. (2022) finds a positive relationship.

[8]For example, see Bayer et al. (2017); Elster and Zussman (2022); Freddie Mac Economic & Housing Research (2021); Kermani and Wong (2021*b*); Kim (2000); Lewis, Emerson and Klineberg (2011); Perry, Rothwell and Harshbarger (2018); Perry (2021); Zhang and Leonard (2021). See Cutler, Glaeser and Vigdor (1999) for a summary of centralized discrimination.

racial wealth gap.

Initially, it was hoped that the use of algorithms would help reduce racial disparities. For example, Kleinberg et al. (2018*b*) show that reliance on algorithms to grant bail could simultaneously reduce crime, jail populations, and racial disparities. However, there are many examples of algorithmic bias, or algorithms that disparately direct fewer opportunities or resources toward minorities.[9] This paper is the first to show the indirect effects of algorithms on racial bias that work via market competition.

Understanding the impact of large investors on the housing market is an important policy question. For instance, in December of 2023, Democrats introduced legislation in the House and Senate that would ban hedge fund ownership of single family homes (Kaysen, 2023; Merkley and Smith, 2023). A growing interdisciplinary body of work has examined the impacts of large single-family investors in the US and, more recently, Europe.[10] This paper speaks to policy discussions around algorithms in the housing market and digitization of public records.

2.1 Background

I provide some background on real estate investment and human and algorithmic investors and elaborate on why real estate investment is fundamentally a prediction task. After setting out the prediction problem, I detail the human and algorithmic investors' approaches to prediction. I describe county real estate records, the timing of the digitization process, and explain why digital county records are significant for algorithmic valuation. Then, I provide some relevant background on racial disparities in the housing market.

2.1.1 Real Estate Investment

Single Family Homes

Residential real estate is the largest asset class in the United States, with a total value of $43 trillion (Malone, 2023). Single-family detached houses comprise 86% of the value of all residential real estate and 66% of the entire housing stock and are common in rural and urban areas (Malone, 2023; Neal, Goodman and Young, 2020). In my sample, single-family houses make up 66% of the occupied housing in urban areas and 72% in rural areas (U.S. Census Bureau, 2021). Single-family houses are purchased by two types of buyers: owner-occupiers, who buy houses to live in, and investors, who buy to rent out or flip these properties.

[9] See Smith (2021) for a summary of empirical work on algorithmic bias. See Rambachan and Roth (2019), Rambachan et al. (2020), Bakalar et al. (2021), Kleinberg, Mullainathan and Raghavan (2016) and Cowgill and Tucker (2019) for theoretical work.

[10] Fields (2018, 2022) examine how technology-driven "calculative agency" enabled the financialization of the single-family housing market. Raymond et al. (2021, 2018, 2016) study the impacts of institutional investors in Georgia and housing insecurity. Mills, Molloy and Zarutskie (2019) provides some empirical early-stage analysis of the activities of these firms. Gurun et al. (2023) study the increase in institutional investor ownership and the impacts of investor mergers on rent and neighborhood safety. Buchak et al. (2022) studies the "i-buyer" firms (e.g. Zillow, Offerpad, Redfin and Opendoor) and their impacts on liquidity in the housing market. Francke et al. (2023) examine the impact of a ban on large institutional buyers of housing in the Netherlands.

Although most strongly associated with homeownership, 17%, or about 14 million of these homes, are occupied by renters. These houses make up the largest single segment of rental housing (about 41%) and are particularly important in areas less urban and with lower income (Census, 2023; Freddie Mac Economic & Housing Research, 2018; Neal, Goodman and Young, 2020).[11]

Investing is a Prediction Problem

Investing depends on the prediction of net income and asset value. Investors forecast possible house income from asset value and rent against upgrades, repairs, and ongoing maintenance. Investors assess the physical condition of the home, both inside and outside, including the structure, the fixtures in the bathroom, and the electrical systems. They will try to confirm actual square footage and configurations; for instance, does the house have an illegal unzoned bedroom that the owner is to claim value for? What do the neighboring houses look like? Does the physical layout of the house make good use of the space? Officially, the metric investors try to predict is the capitalization rate, or net income divided by asset value. The capitalization rate is a standard metric used by investors in the real estate industry to compare properties.[12]

The Human Informational Advantage and "Mom and Pop" Investor

Local entrepreneurs are best positioned to solve this prediction problem. In fact, it was widely believed that these "mom-and-pop" entrepreneurs would always dominate the single family home market due to their informational advantage (American Homes 4 Rent, 2013, 2018; Fields, 2018). Local residents know where traffic is bad, which neighborhoods have the best parks, recent patterns of gentrification, and closings of manufacturing plants or retail stores. These "mom-and-pops" also often had a background in construction and local real estate, helping them evaluate the costs and time required to complete each repair accurately. Individuals familiar with local construction practices can estimate how much exposure to moisture will degrade the foundation of the house. The average mom-and-pop owns a single property in their local area and often works in construction or real estate (American Homes 4 Rent, 2013; Fields, 2022).

In principle, a single company could employ a large number of individuals to evaluate and acquire property. However, evaluating single family homes that are scattered and structurally unique with people is prohibitively expensive in terms of money and time (Amherst, 2016). In the early 2000s, Redbrick Partners attempted to assemble a large portfolio of single-family homes using this strategy. The firm amassed 1,000 homes over the next four years, but struggled to acquire and manage individual houses efficiently. Despite the rapidly increasing price of houses, the firm determined that it was too costly to deal with spatially and physically distinct housing units without technology (Mills, Molloy and Zarutskie, 2019). As a result, Redbrick Partners decided to exit the business in 2006 (Fields, 2018).

[11] In the US, single-family homes are detached dwellings built to be occupied by one household on their own plot of land.

[12] Appendix Figure A.1, shows an example of the capitalization rate information provided for a multifamily property.

Investing is a Challenging Cognitive Task

Although human investors have access to huge amounts of information, processing all of this into a single estimate is a challenging cognitive task. How to weigh the value of going from a one-car garage to two-car garage while taking into account the nice local park, old bathroom fixtures, and trees that may need to be cut down? In scenarios where humans have to weigh a lot of different information, human cognitive limitations can hinder accuracy. For example, Mullainathan and Obermeyer (2021) show that doctors seem to rely on an overly simplistic model to predict heart attacks and overestimate the importance of physical symptoms such as chest pain. Humans also generally do not have experience in learning from thousands of houses—they are limited to their own experience.

Human investors expend substantial effort to structure their decision processes to avoid errors. An industry standard practice is to develop a "buy box" that guides which houses they *will* buy.[13] A buy box is a list of criteria that outline where an investor buys houses and where they believe they have an advantage in valuing houses. An investor might have a buy box that targets houses in a specific zip code in Fresno, California where "the middle class lives" with a diverse employment base between two and three bedrooms. Another industry standard best practice is a blacklist that *eliminates* houses from consideration. For example, an investor could avoid all houses that require electrical system, roof, or septic tank repair because these construction projects are notoriously unpredictable. These tools are efforts to help investors avoid two well-known pitfalls: buying houses that "feel like a great deal" or houses that are aesthetically pleasing but with significant structural flaws. However, doing this well is challenging. Redfin estimates that investors lose money on one in seven homes (Redfin, 2023*a*).

Using Algorithms to Value Houses

ML algorithms, also known as acquisition, automated valuation, or acquisition engines, use statistical patterns in the data to predict the value of the house. Investors use a wide variety of data sources: population, homeownership, vacancy rates, income, crime index, school quality, recent transactions, type of construction, ongoing capital needs, and employment, among others (Amherst, 2016; Invitation Homes, 2017). The set of possible factors that might be useful in estimating net income and asset value is high dimensional. The high dimensional nature of the data creates considerable risk of in-sample overfitting, leading to bad out-of-sample prediction. ML algorithms, which strike a balance between penalizing model complexity and maximizing accuracy, are crucial to doing this well. ML is also necessary because the house value has no explicit formula, requiring a data-driven mechanism to identify patterns and correlations. These algorithms distill the enormous amount of information available into a single estimate of net income. Although building these technology platforms is expensive and requires specialized teams of data scientists and software engineers, "[w]ithout using technology to filter and deliver automated valuations... it would be extremely time-consuming and inefficient to review and bid on these properties... The entire process uses a vast amount of data that is impossible to distill into actionable information

[13]This is also an algorithm, but not a ML algorithm. For example, see New Investors Must Start with a Buy Box or they are wasting time and money.

without the use of technology" (Amherst, 2016; Christophers, 2023).

These algorithms are embedded within an "acquisition team" of human analysts who monitor the houses found by the algorithms. Unlike human investors, these buying teams do not drive around neighborhoods looking for houses. Instead, the algorithms filter through all available houses for sale, estimating net income. The most attractive houses are sent to a queue for the buying teams to review from their desktops (Fields, 2022).[14] Many of the acquisition teams are located in New York, California, and Texas and may never visit the neighborhood where they own houses. The buying team takes the list of properties found by the algorithms, reviews them, and manages the process of generating an offer.[15] Although the algorithm is embedded in a human buying team, the buying team does not physically search for properties themselves and the algorithm does not have access to all of the information available if they were to visit a property in person.[16] I will formalize this trade-off more explicitly in Section 2.4.1.

Algorithms Enable a "Factory-Like" Production Line

ML algorithms produce a single quantifiable house value estimate that can be interpreted without the need for a local context. In the words of one analyst, "... capital markets cannot get into a home... So, [algorithms] take all expenses, all maintenance, water heaters, roof, and flatten them into a format that can be consumed by capital markets" (Fields, 2018). ML algorithms "flatten" a single-family home into a numerical estimate of net income that can be integrated into formal decision-making processes, without the need for deep knowledge of local construction practices. This reduces the cost of acquiring and managing single family homes faced by Redbrick Partners, who found it too costly and inefficient without such a technology.[17] Algorithms helped create a "factory-like" production line for the acquisition, renovation, and leasing of single-family homes (Fields, 2018).

2.1.2 County Housing Records

Algorithms depend on housing data produced by county governments. In this section, I describe county records, the digitization process, and the impacts of digitization on investors in the housing market.

[14] According to its IPO prospectus, Invitation Homes, one of the largest single-family investors, underwrote more than a million homes to assemble its portfolio of 50,000 properties.

[15] Many algorithmic firms employ their own internal real estate agents to make offers on properties. Offers are made primarily by real estate agents to homeowners.

[16] According to Amherst Residential, about five hundred homes newly for sale are listed daily within its target geographic markets and Amherst Explorer, the firm's algorithm filters these listings and delivers automated valuations by estimating potential rents, refurbishing costs, taxes, insurance, and other expenses to calculate an estimated net operating income. Each morning, the firm has a list of targeted properties with projected returns that run automatically before anyone even has had time to drink coffee Amherst (2016).

[17] Furthermore, unlike small investors, who also oversee the property repairs themselves, algorithmic valuation helps other individuals at the same firm make decisions around repair and maintenance costs without ever seeing the house in person.

The Process of County Record Digitization

County governments' records are the most accurate and up-to-date sources of housing market activity and the characteristics of the housing stock.[18] These records, which were kept in paper books or microfilm, are frequently used in day-to-day county business. Dividing property in a divorce proceeding, building and engineering planning, genealogy research, and verifying property ownership all require access. However, paper and microfilm records are not easy to search for, are expensive to maintain, susceptible to physical damage, and are difficult to access and store. Spurred by the clear downsides of paper records and the 2009 Obama Administration efforts to promote digital and transparent government, many counties began to digitize. Digitization transferred these paper and microfilm records into a digital database and made them publicly accessible and searchable on the Internet (The White House, 2009).[19] Panel A of Figure 2.1 shows the share of counties with publicly available digitized records. The share of counties with digitized records rose from 40% to 80% by the mid-2010s.

The time to complete digitization was determined by legislative and budget allocation decisions, as well as by technical difficulties of digitization. First, each state had to ensure that county recorders could legally store their records digitally.[20] Then, each county needed to allocate funding. Digitization required scanning and indexing each paper or microfilm record, a time-consuming and expensive process. Counties tended to work backward from their most recent records, digitizing houses by year of the last sale. Next, each county needed to construct a software database, requiring significant investment in information technology, and connect this database to their website. Finally, each state determined a common standard for computer systems.

The time to complete digitization varied significantly between counties. Panel B of Figure 2.1 shows the share of counties with publicly available digitized recorder systems by state over time. In general, due to the coordination required to digitize records, there are sharp spikes in the share of digitization within each state. However, the year each county process varied significantly due to unexpected issues with setting up the database, digitizing records, or funding, leading to idiosyncratic variation.

How County Digitization Changes the Housing Market

Digitization affects the housing market through three channels: real-time data availability, training data derived from historical transactions, and comprehensive information on the characteristics of each house. Once a county transitions to digital records, all new housing sales become immediately available online. This real-time, reliable, and accurate information is vital in enabling algorithms to update promptly, learning which houses are on sale

[18] By law, County governments are responsible for maintaining public records of property; the Recorder's office maintains and preserves all legal documents affecting title to real property, and the Assessor's office determines the value of real property to collect property taxes. Deed records are public records that date back to county founding; some land records date back to the 1600s.

[19] Because this information is public data, digitizing these records also required making them accessible online.

[20] I use the year county Recorder deed records are first available. In practice, property characteristics data also generally become available at this time.

and which have recently sold. The real-time availability of digital data was consistently highlighted as the most impactful change.[21] Digitization also makes it easy to download historical transaction data. This digital information serves as training data for the algorithms discussed in Section 3.4.

To predict value, algorithms require a digital representation that includes its characteristics—the number of bedrooms, bathrooms, stories, and whether it has a basement. When a house is in the database, digital records of the characteristics of the house are readily available and easy to assess algorithmically. If the house has not yet been added, investors would need to manually collect these data to estimate the value, making it harder to value these houses with an algorithm. [22] I leverage the bureaucratic variation in when each house was digitized to perform robustness checks and estimate house-level effects.

2.1.3 Racial and Ethnic Price Differences in the Housing Market

Prior to the passage of the Fair Housing Act, race was explicitly taken into account when estimating the value of a house. For example, "The Valuation of Real Estate," a popular textbook for real estate appraisal, claimed that neighborhood decline inevitably results from occupation by "...the poorest, most incompetent, and least desirable groups in the city," and described how "... racial heritage and tendencies seem to be of paramount importance" in determining property values (Babcock, 1932; Wheaton, 2023). While it is now illegal to explicitly incorporate race, racial disparities remain in house values. I review the evidence on racial disparities in house values. In Section 2.4.5, I examine racial disparities in my sample.

Evidence on Racial Disparities in House Prices

Racial disparities exist in house prices. Harris (1999) documents that moving from a less than 10% Black to between 10% and 60% Black neighborhood is associated with a 2.3% drop in house value, accounting for house and neighborhood characteristics. Perry, Rothwell and Harshbarger (2018) estimate that homes lose 23% of value when moving from a census tract with 0% Black residents to one that is 50% Black. At the building level, Elster and Zussman (2022) find that house prices decrease 2 to 3% after minorities move in.

Price disparities could reflect preferences, omitted variables, or biases. White homebuyers exhibit a strong negative outgroup bias (negative perceptions or prejudices towards those not in their group) toward living in areas with Black and Hispanic neighbors. Minority home buyers do not show strong preferences and are willing to live in a variety of places, including majority White neighborhoods (Lewis, Emerson and Klineberg, 2011). Minority neighborhoods could be associated with higher crime, lower investment, and lower property values (Freddie Mac Economic & Housing Research, 2021; Harris, 1999; Howell and Korver-Glenn, 2018; Lewis, Emerson and Klineberg, 2011; Perry, 2021). Price differences could reflect omitted variables correlated with the race of the homeowner, such as differences in

[21] Interview with the Georgia Superior Court Clerks' Cooperative Authority. MLS data and Zillow data are considered unreliable because they depend on accurate data entry from real estate agents and are generally not updated in real time. Private data providers, especially in the early 2010s, either did not have data or failed to provide real-time data updates.

[22] Collecting this data by hand is possible, but significantly more costly.

neighborhood amenities or house characteristics. For example, levels of pollution and noise are typically higher in minority neighborhoods (Casey et al., 2017; Tessum et al., 2021).

Yet, neighborhood characteristics cannot fully explain price disparities because disparities persist even when considering the value of the same house. Appraisal is the process through which a real estate appraiser estimates the fair market value of a house for property tax or credit purposes. Widespread anecdotal accounts of appraisal undervaluation have been reported for minority buyers. After receiving a low appraised house value, some minority homeowners have tried to "whitewash" their homes by removing all family photos, asking a White friend to stand in as the homeowner, and received higher estimates of house value in a second appraisal (Howell and Korver-Glenn, 2018; Kamin, 2023; Lilien, 2023). A very small audit study of this found that, on average, a White homeowner received a 7% higher appraisal than a minority couple for the same house (Lilien, 2023). In general, minority-owned homes are more likely to receive appraisal estimates below what a buyer has offered to pay, even when considering the characteristics of the house and the neighborhood (Freddie Mac Economic & Housing Research, 2021; Howell and Korver-Glenn, 2018; Perry, 2021). This suggests that the omitted variables may not fully explain racial disparities in prices. In Section 2.4.5, I examine racial disparities in prices in my data.

2.2 Data and Empirical Strategy

2.2.1 Empirical Strategy: Digitization

My analysis uses a difference-in-difference (DiD) analysis. I use a dynamic event study with differential timing to isolate the causal impact of digitization on the entry of algorithmic investors and market-level and house-level outcomes:

$$y_{ct} = \delta_t + \alpha_c + \sum_{j \neq -1}^{J} \beta^j \times \mathbb{1}[t = j] \times D_{ct} + \gamma X_{ct} + \epsilon_{ct} \tag{2.1}$$

The outcome variables y_{ct} capture the results for county c and year t. First, I examine the impact of digitization on algorithmic investor entry. The outcome is $y_{ct} = ln(1 + q_{ct}^{algo})$, the natural log transformation of one plus the number of houses purchased by algorithmic investors (q_{ct}^{algo}) in county c and year t. I estimate the impact of digitization on price using $y_{ct} = ln(price_{ct})$ or the natural log of the county average sale price of houses in year t. D_{ct} is an indicator equal to one if county c has digitized in year t and 0 otherwise. Digitization is an absorbing state; once a county builds a database system, they do not return to paper records. Counties that had not been digitized by 2017 are used as controls. All regressions include year fixed effects (δ_t) to account for factors that vary over time such as interest rates, housing market policy and other macroeconomic variables. I also account for time-invariant factors specific to each county, such as size, income levels, and geography (α_c). Standard errors are clustered at the county level. The β^j vector is the parameter of interest that captures the time-varying treatment effect of digitization. At the county level, I weight the regressions based on the number of property transactions in each county-year.

I use a series of dynamic differences in difference estimators that are robust to the effects of digitization varying over time. The treatment effects of digitization could increase over

time as algorithms may become more accurate and organizational processes are established. On the other hand, the effects of treatment could also decrease as competition in the housing market intensifies. To address time-varying treatment effects, I use the Sun and Abraham (2021) interaction weighted estimator (IW) that is robust to the correlation over time and across adoption cohorts. I also present results using a series of additional robust estimators introduced by de Chaisemartin and D'Haultfœuille (2020), Borusyak, Jaravel and Spiess (2022), Callaway and Sant'Anna (2021) as well as using traditional two-way fixed effects regression analysis. In general, estimates from robust estimators are larger and more stable because they avoid comparisons between already-treated counties.

These estimators require three assumptions: no anticipation, no spillovers between treated and not-yet-treated counties, and parallel trends. First, participants should not change their behavior in anticipation of future treatment. Second, digitization in one county should not impact the housing market in a county that has not yet been digitized. Third, in the absence of treatment, the treatment and control groups would have evolved similarly.[23] For example, there should be no changes in county economic policy that differentially impact treatment and controls. In Section 2.3.2, I examine the robustness to a series of alternative explanations.

In Figure 2.1, I plot the share of counties with accessible and digitized county Recorder databases over time. The sharp nature of digitization patterns is important to my empirical strategy. The discrete change in digitization will generate discrete changes in algorithm availability, while other unobservables should evolve smoothly around the threshold.

I also estimate a series of cross-sectional hedonic regressions at the house level. This complements the county-level analysis and allows me to explore the impact of house-level digitization (D_{ict}) on house-level outcomes, accounting for differences in observable house characteristics. We examine the likelihood that an algorithmic investor purchases an available house, denoted $\mathbb{1}[q_{ict}^{algo} = 1]$, and a natural logarithmic transformation of the sale price. At the house level, algorithmic purchase could be correlated with unobserved aspects of the house, the number of bidders, time on the market, or the tech-savviness of the listing real estate agent. To address this, I also perform a Two-Stage Least Squares (2SLS) regression, where purchase by an algorithmic investor is instrumented with digitization (Angrist, Imbens and Rubin, 1996).

$$\mathbb{1}[q_{ict}^{algo} = 1] = \delta_t + \alpha_g + \beta D_{ict} + \gamma X_{ict} + \epsilon_{ict} \tag{2.2}$$

The second stage of the relevant house level regression, run using 2SLS to obtain correct standard errors, is:

$$y_{ict} = \delta_t + \alpha_g + \beta \times \mathbb{1}[\widehat{q_{ict}^{algo} = 1}] + \gamma X_{ict} + \epsilon_{ict} \tag{2.3}$$

2.2.2 Data

My sample includes data from 400 counties in Georgia, North Carolina, South Carolina, and Tennessee, spanning the period between 2009 and 2021. Information on property records comes from the county governments. I use detailed property-level house characteristics and

[23] In another way of saying the same thing, the timing of digitization is not correlated with first stage or reduced form outcomes.

sales information from ATTOM Data and Zillow. I also rely on aggregated rental and listing data from Zillow and demographic and socioeconomic data from the US Census.

Digitization Data

I hand-collected data on county record digitization from county recorders' offices, the Internet Archive, and ATTOM Data. The primary source of information was direct interviews with county officials. County officials provided the year when their transaction records first became publicly available online. Once counties switched to electronic records, all future property transactions were automatically digitized, and information on recent transactions became immediately available online. Database systems also enabled easy download of historical data and house information.

I supplemented these interviews with snapshots of county websites from the Internet Archive. These snapshots provide verification of when the county website first provides remote access to the county records. Counties did not keep systematic records when each house record was digitized. Instead, I collect this information from ATTOM Data, who tracked when each record was added. I discuss further details on digitization in Sections 2.1.2 and 2.1.2.

A central concern with hand-collected data is the potential for measurement error. To address this, I use the digitization year provided by ATTOM to corroborate the county information. Although these two series will not align perfectly—-since houses are not all digitized at once and new houses are continually added—the two are similar. To validate the year of digitization of the ATTOM house record, I compared the year provided by ATTOM with the year of digitization from a subset of Georgia counties that maintained more detailed records of house-level digitization. While these records are no longer updated, I collected copies of this information stored by the Internet Archive. Once again, the ATTOM Data year of digitization closely corresponds to county records.

Identifying Investors

Investors are corporate entities that buy houses to rent out or resell homes (Redfin, 2023*b*). I exclude government entities, banks, credit unions, timeshare operators, securitized mortgage trusts, homeowner associations, churches, corporate relocation services, hotels, vacation rentals, farms, builders, and property owner associations. This definition follows other work on investors in the single-family market (Mills, Molloy and Zarutskie, 2019; Redfin, 2023*b*).

After identifying all investors, I categorize each firm as human or algorithmic. I identify algorithmic investors and their properties using business registration information, public filings, and personnel records. I start with properties owned by corporate entities and identify corporate mailing addresses (Gurun et al., 2023; Mills, Molloy and Zarutskie, 2019). To match subsidiaries to the parent firm, I perform two rounds of fuzzy clustering, first on the mailing address and then on public business registration data, properties listed on landlord websites, and known lists of corporate subsidiaries from SEC filings. After this two-round matching procedure, I determine whether each firm's investment strategy is algorithmic or not using SEC filings, news articles and interviews, company websites and personnel records. If the company uses an "algorithmic acquisition engine" or "automated valuation platform" or

employs a data science or software engineering team, I code them as algorithmic. Consistent with previous studies, I find about 40 algorithmic investors in my data, which own about 130,000 houses (Gurun et al., 2023; Mills, Molloy and Zarutskie, 2019).[24] Although not all companies using algorithmic valuation conduct interviews or file with the SEC, all companies have business registration data, websites, and personnel records available on LinkedIn.[25]

I identify human investors as those using non-algorithmic acquisition strategies. I rely on news articles, interviews, company websites, and personnel records to determine whether a firm relies primarily on human judgment to evaluate investments. As a result of the dominance of the mom-and-pop entrepreneur, no human investors are public firms that file with the SEC. However, most of the human investors have websites or personnel records available on LinkedIn, and all have business registration data. Due to the time-intensive nature of this search process, I only explicitly categorize firms with at least 80 purchases in my sample. Of the entities with less than 80 purchases over the decade in my sample, of those that are not categorized as algorithmic, I assume that these are investors using human judgment.

Housing Market Data

Residential housing market comes from ATTOM Data and Zillow's ZTRAX database. Both sources provide records from county Recorder offices and county property tax assessor records. The recorder office data include detailed property transaction information, including sale price, date, identities of buyers and sellers, the corporate structure of the buyer or seller, any relationship between the two, and indicators for arms-length transactions and sales of newly constructed houses.

The tax assessor records provide detailed property and yard characteristics such as property type, longitude and latitude, year built, architectural style, number of bedrooms and bathrooms, type of air conditioning and roof construction material. The records also include estimates of the house market value, land and improvements over time. As of 2023, all housing records in this sample dating back to the early 2000s have been digitized, enabling historical analysis. I drop non-arms-length and multiparcel transactions. I geocode each house to the corresponding census county, tract, block group, and block, using latitude and longitude. Two percent of houses cannot be geocoded to the census block level, but all houses are geocoded to the census block group level.

I supplement the house-level transaction files with various publicly available information from Zillow on housing market dynamics. These measures include the average sale price to the list price, the share of listings with price cuts, the median sale prices, and the share of sales over the list price at the zip code and county level.

In addition to the information on each house, I scraped exterior and interior house images. I collect these images from Zillow and investor websites where these properties are listed. Images are only available for a subset of the houses in my sample, about 50,000 houses. Then,

[24]There are a series of consolidations between the algorithmic investors in the dataset such that at the end of the sample, the total number of firms is smaller.

[25]Only algorithmic investors that are publicly traded REITS or involved sale of securities to investors, must submit SEC filings.

I process exterior house images into vector embeddings for analysis using a deep learning model, which I describe in Section 2.4.5.

I also use a variety of socioeconomic and demographic variables from the American Community Survey (ACS) and the 2010 and 2020 Decennial Census. Many of the counties in my study have fewer than 65,000 residents and thus do not meet the 1-year ACS inclusion threshold, and I rely on the 5-year ACS waves instead. I use a variety of demographic and socioeconomic data, including factors such as median income, median age, racial composition, education, the fraction of the population that is rent burdened, median rent, household size, share in the labor force, share unemployed in the county, census tract, block group, and block level.

Identifying Homeowner Race

I use the Bayesian Improved Surname Geocoding (BISG) proxy method to infer race and ethnicity from publicly available homeowner names. The BISG model predicts race and ethnicity based on owners' surnames and census block addresses using Bayes' theorem. This approach is widely adopted in fair lending analysis (Elliott et al., 2009). The Consumer Financial Protection Bureau, which uses this algorithm for fair lending analysis, has conducted accuracy tests in mortgage lending, a setting that closely mirrors my own (Consumer Financial Protection Bureau, 2023). Using census block geocoding, BISG exhibits Area Under the Curve (AUC) scores of 0.94 or higher across classifications, including Hispanic, Black, non-Hispanic White and Asian borrowers. This suggests the model can accurately categorize races and ethnicity from geography and surname information.[26]

2.2.3 Summary Statistics

I present some summary statistics on the houses purchased by owner-occupiers, human investors, and algorithmic investors.

House-Level Descriptive Statistics

Owner-occupiers make up the bulk of the market. Owner occupiers buy about 86% of all houses as shown in column 1 of Table 2.1. On average, they purchase 2.12 bedroom, 2.14 bathroom houses, 30 year old houses with a garage, a parking space, and a fireplace for about $194,270.

Human investors purchase on average less expensive, older homes. However, these houses are not significantly different from the overall population. Column 2 of Table 2.1 shows that human investors are more likely to buy slightly smaller, less expensive, and older houses around $127,755 that are less likely to have a garage and a parking space.

Algorithmic investors tend to purchase newer, larger, and more expensive homes. As shown in column 3 of Table 2.1, these investors focus on properties with 2.76 bedrooms, 2.47 bathrooms, a mean transaction price of $219,130, and almost always include a parking

[26] AUC scores range from 0 to 1 and represent the model's classification accuracy. A score of 0.5 indicates that the model performs no better than random guessing, while 1 indicates perfect classification.

space. The houses they purchase are, on average, only 21 years old and were remodeled 18 years ago.

The most striking difference between human and algorithmic investors is the very low variation in characteristics of houses purchased by algorithms and the very large standard deviations among houses purchased by human investors. The standard deviation on all house characteristics in column 3 of Table 2.1 are much smaller than column 2. These differences can be illustrated even more clearly in Panels A through D of Appendix Figure A.5. The distribution of houses purchased by algorithmic investors, relative to human investors, is much more concentrated in terms of bedrooms, bathrooms, age, and sale price. I will return to this in more detail in Section 2.4.

Table 2.2 shows the county-level characteristics of the houses purchased by human and algorithmic investors. Algorithmic investors are active in slightly larger and wealthier counties with a higher Hispanic population. Otherwise, the characteristics of the county are relatively similar.

Firm-Level Descriptive Statistics

Algorithmic firms also tend to be much larger firms operating in wide geographic areas. As shown in Panel E of Appendix Figure A.5, before digitization, the market is dominated by many small firms. Conditional on participating in the market, the average human investor purchases a single house. Algorithmic firms purchase an average of 2,000 houses a year. As a result, once a county digitizes, the scale of the largest firms in the market increases significantly. Panels F of Appendix Figure A.5 shows algorithmic investors active in close to 300 different zip codes each year. They have less than 5% of purchases in the same zip code as their corporate mailing address. However, 40% of the houses bought by human investors are in the same code as the investor's corporate mailing address.

2.3 Digitization Leads to Algorithmic Investor Entry

2.3.1 County Digitization and Entry

The raw data clearly demonstrates the impact of digitization on the buying behavior of algorithmic investors. Panel A Figure 2.2 shows the natural log transformation of the number of houses purchased by algorithmic investors in each county, by time to digitization. Panel B of Figure 2.2 shows a sharp increase in market share, which increased from nearly zero before digitization. Following digitization, algorithmic investors buy on average 2% of all houses sold. Panel C of Figure 2.2 shows that the number of houses bought by human investors does not change.

Figure 2.3 presents the analysis of the accompanying event study that shows similar large and persistent increases in the number and share of homes bought by algorithmic investors. Panel A of Figure 2.3 shows that county digitization leads to a 200 log point increase in the number of houses purchased by algorithmic investors following digitization.[27] This increase persists and remains stable until the end of the sample period. Panel B of Figure 2.3 shows

[27]The increase is $e^(2) - 1 = 6.4$

that digitization is associated with an increase in market share of algorithmic investors of 2%. All regressions are adjusted for county- and year-fixed effects and weighted by the number of transactions. Standard errors are clustered at the county level. County- and year-fixed effects account for time-varying common shocks that impact the housing market and county-specific characteristics.

Alternative estimators show similar results. In Appendix Figure A.3, I show the results are similar using alternative event study estimators: Borusyak, Jaravel and Spiess (2022), Sun and Abraham (2021), de Chaisemartin and D'Haultfœuille (2020) and the traditional fixed two-way effects model. Robust estimators avoid comparing newly treated units with already treated units, thus delivering larger and more stable estimates than the two-way fixed effects model.[28]

In Table 2.3, I present the corresponding DiD estimates. Across estimates, I find digitization leads to large increases in the number of houses purchased by algorithmic investors. The Callaway and Sant'Anna (2021) estimates of a 100 log point increase are lower because this estimator cannot be weighted by county size. Taken together, I interpret these results to suggest that county digitization and the subsequent increase in data availability, on average, lead to a sharp and sustained increase in home purchases by algorithmic investors.

The timing of county digitization is not related to observable county characteristics. Table A.1 shows that early and late digitizing counties are balanced in unemployment, income, other demographics, rent, and vacancy rates. Early digitizing counties are larger and have a 1-percent higher Hispanic population than late digitizing counties, but are otherwise similar in socioeconomic and demographics. In Appendix Table A.4, I show how estimates from the standard DiD vary with additional controls. In column 1, I show that, controlling for county and year fixed effects, digitization increases the number of houses purchased by algorithmic investors by 113 log points. In column 2, I shows how estimates vary with additional controls for pre-digitization county socioeconomic status, including demographics, poverty, unemployment, share with young children and educational attainment. In column 3, I add controls for the pre-digitization number of housing units and rent burden. In general, the estimates fall slightly, but remain stable. I interpret these results to suggest that my estimates of the impact of digitization are not driven by systematic differences in observables between counties.

I also see similar strong impacts at the house level. Column 1 of Table 2.4 shows that digitization results in a 17-fold increase in the likelihood that an algorithmic investor purchases a home compared to a non-digitized house in the same census tract. Column 2 indicates a 16-fold increase compared to a non-digitized house in the same census block group. Column 3 demonstrates a 7-fold increase within the same census block. These results suggest that algorithmic investors are sensitive to the availability of digital information when valuing houses.

[28] Callaway and Sant'Anna (2021) cannot be weighted with the number of transactions, so I only plot the other estimators.

2.3.2 Within County Triple Difference and Falsification Tests

I next address if there are unobservable factors that affect both algorithmic investor activity and the timing of digitization. For instance, county officials might be working to attract business investment and modernize government processes. To investigate this, I leverage house-level variation in the cost of algorithmically valuing houses to test for evidence of unobserved shocks.

The timing of house-level digitization is not related to house or neighborhood attributes. Appendix Table A.2 reveals that the early and late digitized houses are evenly matched in features such as the number of bedrooms and bathrooms and the presence of a basement or other structures.[29] While houses that are digitized later show a more recent last sale date, newly constructed houses are also digitized later such that there are no substantial disparities. Appendix Table A.3 shows that houses are also similar on neighborhood characteristics.

To investigate common county shocks, I compare digitized and not-yet-digitized houses within the same county before and after digitization. Panel A of Figure 2.4 plots the raw data, showing the number of houses bought by algorithmic investors separately for digitized and non-digitized houses. Algorithmic investors mostly purchase houses that exist in the county's database. These investors buy very small numbers of non-digitized houses. These outcomes could potentially be attributed to measurement errors in county record keeping, misclassification of algorithmic investors, transactions involving the purchase of multiple houses, or scenarios where algorithmic investors supplement county databases with additional data.

Panel B of Figure 2.4 displays the corresponding interaction-weighted event study at the county level separately for digitized and non-digitized houses. This panel illustrates that the increase in the number of homes purchased by algorithmic investors is almost entirely confined to houses with digital records. Unobserved county-level shocks, such as changes in housing or foreclosure policy, should impact all houses in a county, regardless of digitization status. County shocks are not consistent with an impact that is so concentrated in digitized houses.

I perform a falsification exercise to assess if county digitization impacts nondigitized houses after adjusting for house-level characteristics, using the regression in Equation 2.4 run at the census block group level. In the equation, β^{Digit} captures the impact of market digitization on algorithmic investor purchases of digitized houses. $\beta^{NoDigit}$ measures the impacts on non-digitized houses. I also include controls for neighborhood and house characteristics.

$$\begin{aligned} 1[q_{igt}^{algo} = 1] = \delta_t + \alpha_g + \beta^{Digit} \times \mathbb{1}[HouseDigitized_{ict} = 1] \times CountyDigit_{ct} + \\ \beta^{NoDigit} \times \mathbb{1}[HouseDigitized_{ict} = 0] \times CountyDigit_{ct} + \gamma X_{igt} + \epsilon_{igt} \end{aligned} \tag{2.4}$$

Column 4 of Table 2.4, shows that county digitization does not impact non-digitized houses; the impact is solely on digitized houses. These results are not consistent with unobserved, neighborhood-level shocks driving investor activity.

However, suppose that the existence in the county database simplifies the house discovery process for *all* investors. Human investors should then also be more likely to buy digitized

[29] For this analysis, "early digitized" refers to those digitized before the county's median digitization year, and "late digitized" were digitized after.

houses. I test this by examining the effect of house-level digitization on the propensity to purchase by individual investors in column 5 of Table 2.4. $\beta^{NoDigit} = 0.0017$ and $\beta^{Digit} = -0.0699$. Digitization reduces the probability of human investment purchase and has no impact on non-digitized houses. I interpret these results to show that digitization affects algorithmic investors differently than human investors.

Together, these results build confidence that digitization and changing data availability drive algorithmic activity. First, if algorithmic investors had some influence on the digitization process, early digitized houses might look different from those that are digitized later. Second, if algorithmic investors were not relying on algorithms to purchase houses, we would not expect their purchases to be so heavily concentrated in digitized houses. Third, if localized neighborhood shocks were driving our results, we would expect both digitized and non-digitized houses in the same local area to be impacted in a similar manner. Lastly, I show that the impact of digitization on the likelihood of purchase is specific to algorithmic investors. Thus, the evidence suggests that algorithmic investment is indeed driven by changes in data availability due to digitization.

2.4 Allocation and Specialization

In this section, I consider the possibility that algorithms and humans have distinct comparative advantages in prediction problems. I begin with a conceptual framework that illustrates the trade-off between humans and machines.

2.4.1 Conceptual Framework

Houses are characterized by an observable X and an unobserved Z, seen by humans, and a common value component Y. Although the underlying data is multidimensional, I will use two unidimensional variables $x(X) = E[Y|X]$ and $z(X, Z) = E[Y|X, Z] - E[Y|X]$ and $E[y|X, Z] = E[y|x, z] = x + z$.

Human investors generate a prediction using both x and z, but may be biased ($\delta(x, z) \geq 0$). Humans may be biased on some houses, but not on others, or may not make systematic errors. For instance, humans seem to overvalue houses with pools and air conditioning during warm weather (Busse et al., 2012).

$$h(x, z) = E[y|x, z] + \delta(x, z)$$

ML-powered algorithms use patterns in data to make predictions. Algorithms look for patterns in thousands of examples, rather than just being limited to their own experience. They are not subject to the same cognitive limitations as humans. For example, algorithms can quantify the specific value of a two-car garage versus a one-car garage, which is likely outside the scope of most humans. Algorithms are also not explicitly impact by factors like warm weather, how they are feeling at the time and prejudices. However, an algorithm cannot see z.

$$m(x) = E[y|x]$$

For any given house, would an algorithm do better or would a human do better at predicting true y? Given a house with true value y, if $|E[y|x] - E[y|x, z]| >> 0$ or $|E[y|x] -$

$y| >> 0$, then the human-accessible private information is important and a human might do better. For instance, some houses are architecturally complex, and where the information available to an algorithm might not capture the house's aesthetic appeal. Or, a house might have a beautiful view of a nearby farm, such that the value of the house and land is higher than would be predicted by an algorithm. However, proximity to a farm can also have several drawbacks: loud mechanical noises, smell of manure, proximity to pesticides, and a higher than usual number of rodents. These factors may lower the house value compared to what would otherwise be predicted by an algorithm. A human can also walk through the interior of a house, estimating repairs and maintenance costs. However, if $\delta(x, z) > 0$, or human decision making is systematically biased, the value of an algorithm might outweigh the importance of private information.

$$m(x) - h(x, z) = \underbrace{E[y|x] - E[y|x, z]}_{\text{informational advantage, } \mu} - \underbrace{\delta(x, z)}_{\text{human bias, } \delta} \tag{2.5}$$

Equation 2.5 highlights the tradeoff between human and algorithmic valuation. if private information z is important, the human informational advantage can outweigh human error. If $\delta(x, z) = 0$ or humans are not biased, humans will do better. If humans make mistakes, the benefits of an algorithm can outweigh the importance of private information.

2.4.2 Measuring House Predictability

Before showing how human investors respond to digitization, I categorize houses by their degree of algorithmic predictability. I construct this measure of how difficult it is to algorithmically predict each house from commonly available observables. I refer to the difficulty of predicting a house from observables as *predictability*.

To construct this, I use the Extreme Gradient Boosting (XGBoost) algorithm predicting the transaction price (Chen and Guestrin, 2016). Given the high-dimensionality of the data and significance of nonlinear relationships, nonparametric models outperform linear models when modeling houses. For example, even a slight increase in square footage could have a significant impact on price in a densely populated neighborhood, while the same would not be true in a rural area. In these cases, nonparametric models, such as tree-based algorithms, are able to capture nuanced, nonlinear relationships, particularly among the many variables that can influence house pricing – location, size, design, age, local amenities, etc.

The XGBoost algorithm operates on a gradient boosting framework in which new models are generated to correct the errors of pre-existing ones. In essence, it creates a robust overall model by combining multiple weak models to improve the accuracy of the prediction according to the regularized objective shown in Equation 2.6. l is the differentiable convex loss function, T is the number of leaves in each tree and w is the leaf weights (Chen and Guestrin, 2016). Intuitively, this objective function balances training loss $l(\hat{y}_i, y_i)$ with L1 regularization (γT) and L2 regularization ($\lambda||w||^2$) components, encouraging both simpler and more generalizable models.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \gamma T + \frac{1}{2}\lambda||w||^2 \tag{2.6}$$

The model is built using pre-digitization data for each county to exclude any impacts from algorithmic investors. I randomly split the data into a training set and 25% held-out test set. Using the training data set, I perform a grid search through the XGBoost hyperparameter space, using 3-fold cross validation with early stopping (LaValle, Branicky and Lindemann, 2004; Shen, Gao and Ma, 2022).

In panel A of Figure 2.5, I plot the predicted versus actual log price for the held-out sample. The average out-of-sample root mean squared error is 0.903. More intepretably, 40% of the houses in the test set are within 10% of the price. The same measures computed for Zillow's Zestimate, which incorporates demand information from user interactions with their website, reveal that 59% of houses are priced within 10% of the sales price in areas with Zillow coverage (Zillow, 2023).

For each house, I calculate the out-of-sample average—the difference between the actual and predicted price—to capture how easy or hard it is to predict each house. The variation in prediction error is enormous, with the average model varying widely: for some houses, it is close to 50%, while for others it is close to zero. Houses in neighborhoods built by the same builder and in the same style are easier to model, but older houses, built before the introduction of modern building codes and full of architectural distinction, are much less standardized. Features such as sentimental, aesthetic, or historical value or proximity to a noisy highway or pungent agricultural property may also play a role. If qualities that are hard to quantify empirically significantly influence a house's value, algorithmic prediction will be less accurate.[30]

Counties also differ in institutional processes for the collection and quality of their housing data. Counties vary in the frequency with which they update their housing records, the quality of their data control, and the thoroughness of the information they collect on each home or transaction. Together, the less informative or accurate the observable information, the more important human private information becomes.

2.4.3 Human Investor Shift Towards Hard to Predict Houses

Human investors purchase houses across the entire distribution of model error; in some instances, the predicted price is far lower than the actual price, while in other cases the predicted price significantly exceeds the actual price. In Panel A of Figure 2.5, I show the predicted versus actual prices of the gradient-boosted tree model described in Section 4.2. In general, the model is best at predicting houses in the middle of the distribution. In Panel B of Figure 2.5, I plot the actual price versus the predicted price for houses in a held-out test set from 2012 and 2013. The houses are colored in light blue if these houses will be purchased by human investors in the future, and houses in purple are those that will be purchased by algorithmic investors. Human investors purchase houses across the entire distribution of model error; in some instances, the predicted price is far lower than the actual price, while in other cases the predicted price significantly exceeds the actual price. While some of these may reflect poor human decision-making, on average, large differences between predicted and actual prices may reflect private information. Unlike human investors, algorithmic investors only purchase houses where the model-predicted price closely approximates the actual price.

[30]Price someone is willing to pay may also depend on mood, weather, or noise.

In other words, they buy houses where the scope for adverse selection is small.

In Figure 2.6, I show how human investors react to digitization. Human investors become 50% less likely to buy houses in the lowest decile of model error, where algorithms are most effective. They become almost twice as likely to buy houses that are most difficult to predict.[31] Human investors become less likely to purchase houses in in deciles 1 to decile 7, and more likely to purchase houses in the top two deciles of average model error. These results are consistent with human investors specializing where human comparative advantage is highest.

2.4.4 Discontinuities: Data errors, zoning rules and lead paint

A testable implication of my conceptual framework is that characteristics that increase the importance of private information, should limit algorithmic investor buying. I investigate this with three specific examples and show that algorithmic investors avoid houses where institutional factors limit algorithm accuracy while human investors do not. These results provide additional evidence that algorithmic investors depend on quantifiable information in the dataset while humans are not so constrained. These results are further evidence for patterns of distinct comparative advantage.

Zoning Rules

I illustrate this first with unusual zoning rules in Wilson County, Tennessee. In Panel A of Figure 2.5, there is a distinct group of houses in Wilson County where the predicted model price is much higher than the actual price. This is a result of unusual zoning rules for bedrooms in Wilson County, which make it difficult to interpret the number of bedrooms in county data. Wilson County only considered a room a legally zoned bedroom if the room also included a specific type of closet. As a result, tax assessor records list most houses as having zero bedrooms, although the "true" number of bedrooms is higher. Algorithms cannot accurately value houses without access to the true number of bedrooms, and the average algorithm error is large in this county. Although Wilson County is close to Nashville and similar to other places where algorithms buy many houses, algorithmic investment in the county is limited. However, human investors invest heavily.

Lead Paint

Houses constructed before lead paint was banned are more difficult to value algorithmically. In the early 1900s, lead was a commonly used additive in paint and other building materials. During the 1960s and 1970s, detailed studies on the effects of lead poisoning led to concerns about health effects in residential structures. The Consumer Product Safety Commission banned lead paint in residential construction in 1978 (The Department of Housing and Urban Development, 2023). Houses built before 1978 may have lead paint, whereas those built after 1978 do not have lead paint. Human investors, who can physically inspect houses, can determine whether lead is a concern and accurately forecast the additional costs necessary to deal with lead exposure, regardless of when a house was built. Algorithmic investors face

[31] In the pre period, likelihood of purchase by an human investor is .12.

uncertainty about lead exposure and face unpredictably higher construction costs to deal with lead exposure in houses built before 1978.[32]

In panel A of Figure 2.7, I test for a discontinuity in the density of houses bought by algorithmic investors, using a local polynomial density estimator (Cattaneo, Jansson and Ma, 2018, 2019). As seen visually in Panel A, the null hypothesis of no discontinuity around 1979 is rejected, with a p-value 0.000. In Panel B, I plot the density of houses purchased by human investors. In this case, with a p-value of 0.295, the null hypothesis that the density shows no evidence of manipulation cannot be rejected. I interpret these results to suggest that algorithmic investors appear to respond to the imposition of lead paint, while human investors do not.

Data Errors

Data entry errors limit the effectiveness of algorithms. There are almost 220,000 houses in the county database dataset with more than 15 bedrooms or 15 bathrooms. These houses reflect data entry errors. Algorithms struggle to accurately value houses with data entry errors because the number of bedrooms and bathrooms is such a crucial piece of information.[33] Panel C of Figure 2.7 shows the number of houses with data errors sold over time; the series is relatively spiky but without any clear trends, indicating that the availability of houses with data errors does not strongly vary over time. Panel D of Figure 2.7 shows the natural log transformation of the number of houses with data errors purchased by human and algorithmic investors. Algorithmic investors avoid purchasing houses with data errors, whereas human investors do not. Data entry errors do not pose problems for human investors who do not rely exclusively on hard information.

In all of these instances, institutional details produce variation in the value of private information and create opportunities for human investors. This highlights how institutional details can shape the effectiveness of algorithms use and create opportunities for human judgment.

2.4.5 Algorithmic Investors Specialize in Minority-Owned Homes

After illustrating human comparative advantage and where human investors focus their efforts after digitization, I now turn to algorithmic investors. I first establish the existence and robustness of a race penalty, suggesting the possibility of human bias, and then show that algorithmic investors disproportionately buy minority-owned homes.

The Race Penalty

Prior to county digitization, I calculate a 5% race penalty—lower sale price received by a minority homeowner compared to a White homeowner—accounting for house and neighbor-

[32]Lead-paint remains the most significant source of lead exposure in the US because many houses were built before 1978 (US EPA, 2014). Any renovation, repair or painting project in a pre-1978 home can easily create dangerous lead dust, requiring special lead-safe contracting procedures and contractors (US EPA, 2013).

[33]In principle, they could collect this information manually, but algorithmic firms are not organizationally set up to do this.

hood characteristics (Bayer et al., 2017; Elster and Zussman, 2022; Freddie Mac Economic & Housing Research, 2021; Kermani and Wong, 2021*b*; Kim, 2000; Lewis, Emerson and Klineberg, 2011; Perry, Rothwell and Harshbarger, 2018; Perry, 2021; Quillian, Lee and Honoré, 2020; Zhang and Leonard, 2021). In Figure 2.8, I plot the race penalty in period *prior* to digitization, controlling for various levels of neighborhood characteristics and observable characteristics of the house. All specifications include year and geography fixed effects.[34]

The first bar in Figure 2.8 shows that minority homes sell at a 14% discount relative to White homes in the same county, adjusting for house characteristics. This gap drops to 7% when adding census tract fixed effects. The 50% decrease in the race penalty suggests that there is a lot of unobserved heterogeneity between houses in the same county.[35] At the census block group level, the implied race penalty is 5%. At the census block level, minority-owned homes sell for about 3% or $4,700 less. All of these numbers are calculated in the years before digitization.

Deep Learning Image Analysis

House characteristics do not fully capture many differences between houses. Houses are structurally unique three-dimensional objects that derive their value from the color of the paint, the landscape, the maintenance, and the cleanliness of the windows. Two houses in the same neighborhood may have completely different architectural styles or states of disrepair. (Choi et al., 2019; Harris, 1999; Pinto and Peter, 2021). Minority homeowners are less wealthy and may invest less in house maintenance and aesthetics (Harris, 1999). The race penalty could simply reflect these differences in house appearance.

I use a deep learning model to calculate the race penalty adjusting for house images. I scraped house images from Zillow and other apartment rental websites. Appendix Figure A.6 shows an example house image. Images are not available for all houses in my sample. I rely on images for a total of 50,000 houses. I use AutoGluon, a deep learning model designed for unstructured data such as images, to convert each exterior image into a high-dimensional embedding vector (Erickson et al., 2020). The position of each image within this vector space is indicative of its visual features or content, ensuring that similar images are close to each other in the embedding space. Adding these deep learning embeddings to the race gap regression will control for previously omitted variables the aesthetic features of the house and the yard.

Race Penalty with Deep Learning

Incorporating house exterior images does not significantly change the race penalty. In Table 2.5, I show the race penalty coefficients, controlling for the quality and appearance of the house with deep learning-generated embeddings. These race penalty estimates are similar to the estimates from Section 2.4.5. For example, at the census block level, the race penalty

[34] Include year by geography fixed effects yields very similar results.

[35] In our sample, census tracts encompass 4,517 people or 2,006 housing units. Census block groups average around 1,610 people in our sample or 716 housing units. A census block contains around 65 people. I used the 2010 population to calculate these averages.

is 2.1% with the image emeddings and 3.3% without images, including block by year fixed effects. The existence and persistence of this race penalty suggest that the race penalty coefficient may reflect more than simply differences in house quality. For instance, consistent with other evidence that individuals associate lower values with the same home when they perceive it to be owned by a minority, humans could be undervaluing minority-owned homes (Lilien, 2023).

Algorithmic Investors buy Minority-Owned Homes

Algorithmic investors disproportionately buy minority-owned homes. As shown in Table 2.6, the impact of digitization of house records is twice as strong for minority homeowners than for White homeowners. In column 1 of Table 2.6, the impact of digitization on a minority-owned home is twice as large relative to a White-owned home in the same census tract or census block group. However, the impact of digitization is six times as large compared to a White-owned house in the same census block. These results suggest that algorithmic investors do not just focus on minority neighborhoods, rather, they specifically target minority houses.

These results are not simply driven by all investors targeting minority-owned homes. In Appendix Table A.5, I demonstrate the effect of digitization by homeowner race within a sample exclusively consisting of investor transactions (human and algorithm), excluding the owner-occupiers (those buying houses to live in). These regressions will illustrate the impact of digitization on the likelihood of purchase by algorithmic investors compared to human investors. Column 1 of Appendix Table A.5 reveals that the effect of digitization on minority-owned homes is five times stronger than on White-owned homes. The impact of digitization is five times as large at the census block group level (column 2), and nine times larger at the census block level (column 3). These results suggest that algorithmic investors are disproportionately likely to buy minority-owned homes, even compared to human investors.

2.5 Prices, Spillovers and Racial Disparities

In this section, I explore the consequences for market-level prices and racial inequalities.

2.5.1 Digitization Shrinks the Race Penalty

First, I explore how digitization affects the race penalty. Panel A of Figure 2.9 plots the race penalty coefficient by time to digitization, including census block group controls.[36] In the year following digitization, the race penalty shrinks from 8% to 4%. The race penalty continues to fall until it disappeared six years after digitization.

In Panel B of Figure 2.9, I investigate the mechanism behind this change. The first two bars in panel B of Figure 2.9 plot the race penalty for purchases of owner-occupiers, those who buy houses to live in, and human investors, prior to digitization. Both pay approximately 5% less for the observably similar house in the same census block group with a minority owner compared to a White homeowner. However, as shown in the blue bar, algorithmic

[36]Digitization varies at the county by year level, so we cannot include geography by year controls.

investors do not exhibit any race penalty. Algorithmic investors pay the same price for an observably similar house regardless of the race of the homeowner.

Interestingly, digitization also reduces the race penalty among owner-occupiers and human investors. The fourth bar in Panel B of Figure 2.9 shows that, after digitization, owner occupiers pay only about 3% less for minority-owned homes. The last bar in Panel B of Figure 2.9 plots the post digitization race penalty for human investors. After digitization, human investors pay only about 1.5% less for minority-owned homes. In Appendix Table A.6, I use a 2SLS analysis is used to address endogeneity concerns around other factors related to bidding behavior that could drive these results. The results are much noisier, but qualitatively similar.

These indirect effects are driven by two mechanisms; price anchoring given higher prices of similar houses and algorithmic investors bidding up the prices of minority homes, even in cases where they do not ultimately acquire the house. Real estate agents set listing prices based on similar, recently sold properties. If properties in minority neighborhoods are priced higher, other houses will have higher listing prices, which subsequently leads to higher sales prices. Furthermore, when algorithmic investors try to buy homes, they can drive up the final price for other buyers. In fact, the share of owner-occupiers is so large that the majority of the impact of digitization on the race penalty is due to these indirect effects.

As a result, Figure 2.10 shows that digitization leads to a 5% increase in the prices of minority homes. Among White homeowners, it is possible that algorithms may not raise prices. If homeowners are willing to sell their homes at a discount in exchange for a prompt offer, could lead to a decline in prices. Instead, we also see an increase; digitization leads to a 1.5% increase in the average sale price of White-owned homes.

2.5.2 Adverse Selection or Human Error?

A natural question is whether algorithms are taking advantage of human mistakes or simply overpaying for unobservably worse homes. Adverse selection has been widely cited as a barrier to the use of algorithms in the housing market and has been widely discussed as the reason why Zillow, an algorithmic investor, decided to stop buying houses (Economist, 2021).

I disentangle adverse selection from human error with two complementary approaches. In the first, I calculate the gross margin on each house sold: the difference between the resale and transaction price. If algorithmic firms overvalue minority homes relative to White-owned homes, then the gross margin on minority homes should be lower compared to White-owned homes. I also calculate the gross margin with the estimates of the house market value from tax assessors. Unlike resale price, which is only available for resold homes, these estimates are available for all homes. However, these are estimates made by the human tax assessor rather than actual transaction prices.[37]

Using my two measures of gross margin, I estimate the following regression for house i bought in year t, resold/assessed in year r in census block c, including purchase year by

[37]Note that if tax assessor evaluations are also biased, our results will be more conservative.

census tract, block group or block and resale or assessment year fixed effects:[38]

$$log(price_{irc}^{resale}) - log(price_{itc}^{sale}) = \gamma X_{irtc} + \beta_1^{algo} \times SellerMinority_{itrc} + \epsilon_{itrc} \quad (2.7)$$

The coefficient β_1 indicates if the margin is systematically different for minority homeowners. If houses purchased from minority homeowners are adversely selected, the margin should be lower, or $\beta_1^{algo} < 0$. However, if the higher prices paid by algorithmic investors for minority homes reflect the true value, then $\beta_1^{algo} \approx 0$ or $\beta_1^{algo} > 0$.

I find no significant differences in the gross margin by race of the homeowner. Columns 1 through 3 of Table 2.7 include census tract, block group, and block by year fixed effects, respectively, among houses bought by algorithmic investors.[39] Columns 1 through 3 of Appendix Table A.7 show similar results using the assessment margin. Among houses purchased by algorithmic investors, the margin on minority-owned homes is not systematically different from that on White-owned homes. In column 4 of Table 2.7 and column 4 of Appendix Table A.7, I explore whether the resale margin differs for homes bought in neighborhoods with greater minority shares. In both cases, I find no strong relationship. These results suggest that the gross margin of the algorithmic investor does not vary with the composition of the neighborhood.

Next, I show that the gross margin on minority-owned investors is *higher* that that on White homes. If minority-owned homes are priced too low, the gross margin should be higher due to the discounted acquisition price. Columns 5 through 7 in Table 2.7 show that the gross margin on minority homes is 10% higher among purchases by human investors. In column 8, I explore whether the margin varies by neighborhood composition. The margin may be higher in more minority neighborhoods that contain a higher share of minority residents, but the estimate is noisy. Column 8 of Appendix Table A.7 shows that the assessment margin is 9% higher in neighborhoods with a higher share of minority residents.

If these results are due to racial preferences or heuristics, the differences may be more pronounced in neighborhoods with more minority residents, where humans may have more trouble accurately valuing houses or biases. These results suggest that the higher prices algorithmic investors pay for minority-owned homes are not driven by adverse selection and may, in fact, reflect algorithm comparative advantage in valuing houses where human biases, prejudices, or cognitive limitations may cloud judgment.

These results are also not consistent with adverse selection among algorithmic firms. If minority-owned houses are unobservably worse, humans should not be willing to pay more for these houses after digitization. Humans can access unosbervable aspects of house quality that are not apparent to algorithms and should not be subject to the same adverse selection concerns.

2.5.3 Spillovers

Thus far, my analysis has focused on the prices of sold homes. However, house values of occupied houses are a key driver of household wealth and play an important role in

[38]The time between purchase and resale or assessment is a linear combination of purchase and resale year, so this would drop from any regression.

[39]Not all houses can be geocoded to a census Block level, but I show all three results.

credit markets (Guren et al., 2020). If minority-owned houses purchased by algorithmic investors are predominantly located in majority White areas and not structurally similar to other minority-owned houses, algorithmic investor activity would not necessarily have spillover effects on unsold minority homes. However, if algorithmic purchases are similar to other minority-owned unsold houses, their activity could also have large indirect impacts on minority homeowner house values and household wealth.

Following Hirano, Imbens and Ridder (2003), using the estimate of the expected price of sold minority homes, $E[P|S=1]$, I write the inverse propensity weighted unsold minority-owned homes house price impact $E[P|U=1]$ as:

$$\begin{aligned}
E[P|U=1] &= \sum_X p(X|U=1)E[P|U=1,X] \\
&= \sum_X \frac{p(U=1|X)p(X)}{p(U=1)}E[P|U=1,X] \\
&= \frac{1}{p(U=1)}\sum_X p(U=1|X)p(X)E[P|U=1,X]\frac{p(X|S=1)p(S=1)}{p(S=1|X)p(X)} \\
&= \frac{p(S=1)}{p(U=1)}\sum_X E[P|S=1,X]\frac{p(U=1|X)p(X|S=1)}{p(S=1|X)} \\
&= \frac{p(S=1)}{p(U=1)}E\left[\frac{p(U=1|X)}{p(S=1|X)}P|S=1\right] \qquad (2.8)
\end{aligned}$$

Equation (3.2) says that I can recover the average impact on the price of minority-owned homes by reweighting the prices of sold minority-owned homes, using a ratio of propensity scores to account for differences in house characteristics. Re-weighting based on the characteristics of the observable houses and the census block, I find an average increase of 6% in the value of minority homes. It is important to emphasize that this analysis relies on a selection on observables assumption when reweighting. Although algorithmic investors may not be able to see unobservable characteristics, part of the impact comes from human investors and owner-occupiers, who have access to unobserved information. If unsold minority houses are very different on unobservables than sold minority houses, this estimate may overstate the impacts.

2.6 Conclusion

Progress in ML and the widespread availability of digitized data opens up a wide set of economic possibilities. This work illustrates how the availability of algorithmic prediction not only influences individual decisions, but also precipitates a range of changes at the market level, affecting participation, firm organization, and equilibrium outcomes. In the housing market, the availability of machine-generated predictions leads to new entrants using algorithms to value houses. Human investors react by moving towards parts of the market where algorithms are least effective. Algorithmic investors buy disproportionately where human decisions are biased, causing large price increases. Six years after digitization, the race penalty disappears. Much of these impacts are due to indirect effects of algorithmic

investors that manifest through the nature of competition. These findings suggest numerous avenues for future research.

First, when algorithms and humans disagree, we cannot assume that the algorithm is correct: unobserved information can lead to algorithm errors. At the same time, we cannot assume that the human is always correct. Instead, the value of the tradeoff depends on the importance of private information and degree of bias in human decisions. A growing number of papers show that human errors can be sufficiently systematic to outweigh the value of private information (Kahneman, Sibony and Sunstein, 2021; Kleinberg et al., 2017*b*; Mullainathan and Obermeyer, 2019; Rambachan, 2022). More work is needed to better understand how the value of this tradeoff varies across people and prediction problems.

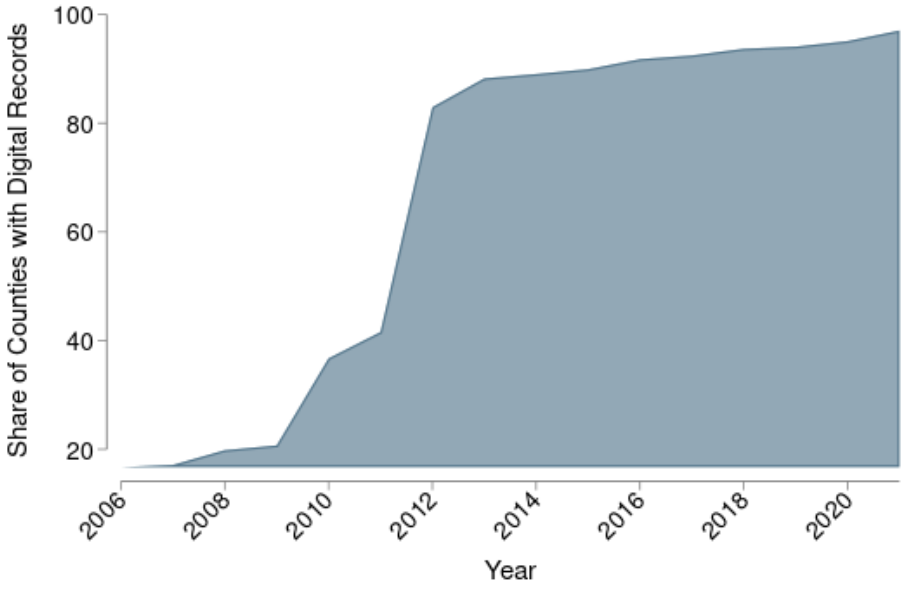
Second, as an evolving technology, the ML tools used by companies employing algorithmic prediction strategies are rapidly changing. In this setting, there is an ecosystem of companies attempting to curate detailed and increasingly accurate datasets, from comprehensive house surveys that measure construction quality to mobile phone data that track neighborhood activities. As data quality improves, the percentage of "predictable" houses that can be targeted for purchase by algorithmic investors may increase.

Furthermore, the efficacy of algorithmic prediction may depend on the specific legal and institutional structures, which vary widely between states. This study examines Georgia, North Carolina, South Carolina, and Tennessee, where housing market transactions and prices are part of the public record. However, in twelve US states, property transaction prices are not automatically included in the public record, thus potentially curtailing the effectiveness of algorithmic prediction. More work is needed to explore how institutional structures impact the potential effects of algorithms.

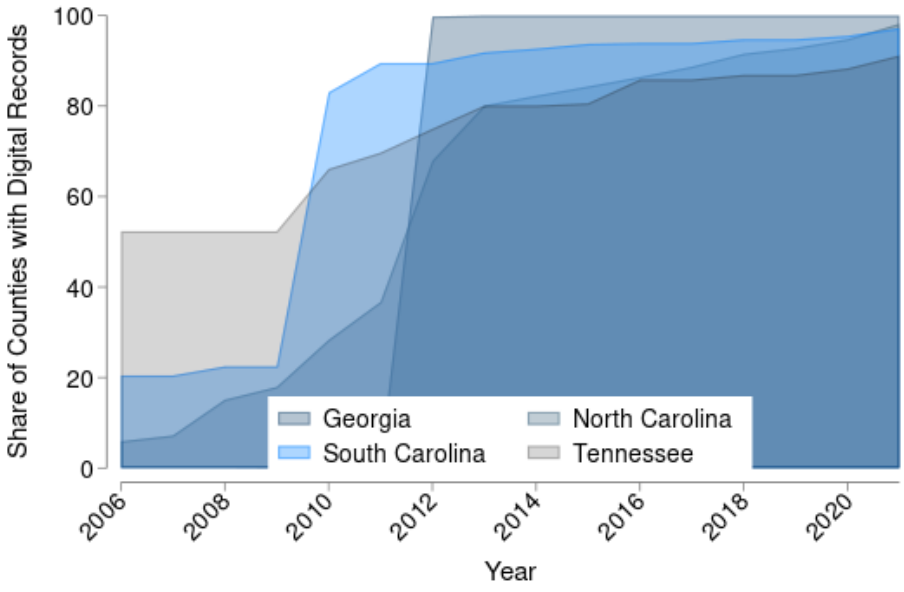
Finally, this study does not dwell on the potential implications of organizational differences between algorithmic and human investors. Algorithmic investors typically operate as large, formal, arms-length organizations, while human investors often manage their rental properties more informally. Unlike human investors, who frequently choose tenants personally or through their social networks, algorithmic firms may rely more heavily on automated screening procedures for tenant selection. Together, these changes could have lasting effects on the local labor market and communities. Given the rapid adoption of algorithms, these effects deserve further study.

Figure 2.1: County Record Digitization

A. Share of Counties with Digitized Records



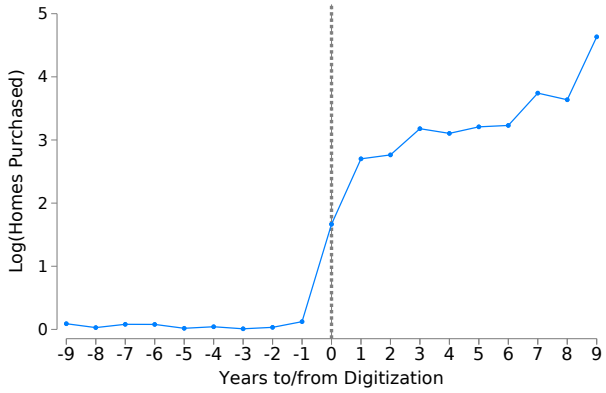
B. Share of Counties with Digitized Records, by State



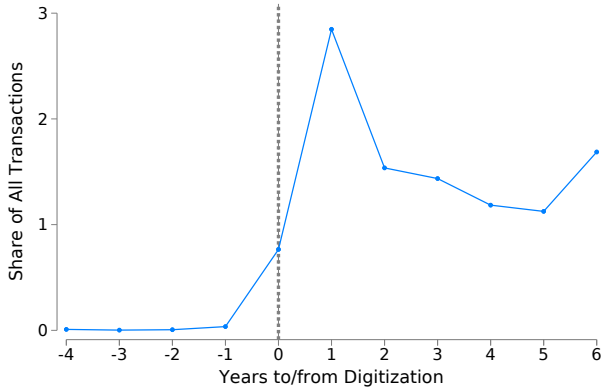
Notes: This figure shows the share of counties in the sample with digitized and publicly accessible Recorder data over time. Panel B shows the share by state. The graphs are weighted by the number of housing transactions. All data comes from county governments.

FIGURE 2.2: ALGORITHMIC INVESTORS BUYING, BY TIME TO DIGITIZATION

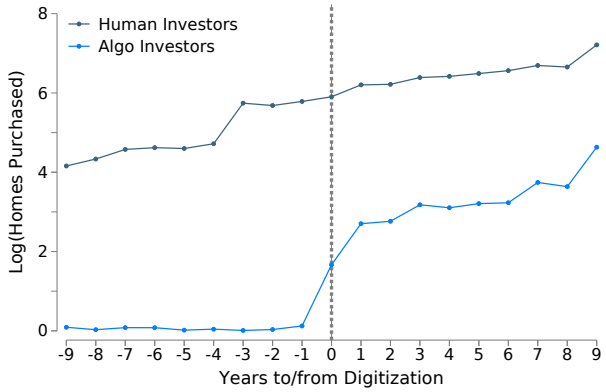
A. LOG(HOUSES PURCHASED BY ALGORITHMIC INVESTORS)



B. ALGORITHMIC INVESTOR SHARE



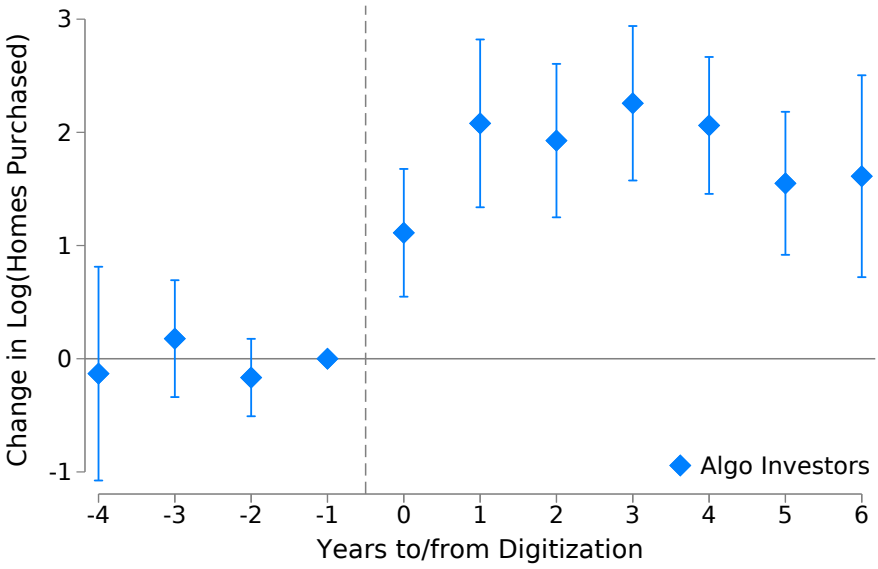
C. LOG(HOUSES PURCHASED)



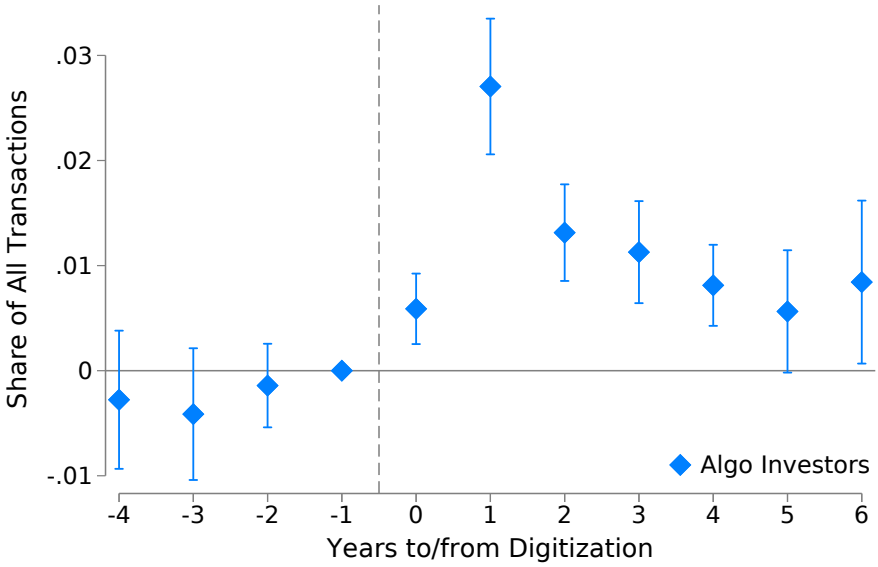
NOTES: This figure shows the number of houses purchased by algorithmic investors in county c and in year t, by time to digitization. Panel A shows the natural log of the number of homes purchased by algorithmic investors. Panel B plots the number of algorithmic investors purchases as a share of all transactions. Panel C adds the natural log of the number of houses purchased by human investors. All data come from ATTOM Data, Zillow and county digitization records.

FIGURE 2.3: EVENT STUDIES, LOG(HOUSES PURCHASED) BY ALGORITHMIC INVESTORS

A. LOG(HOUSES PURCHASED)



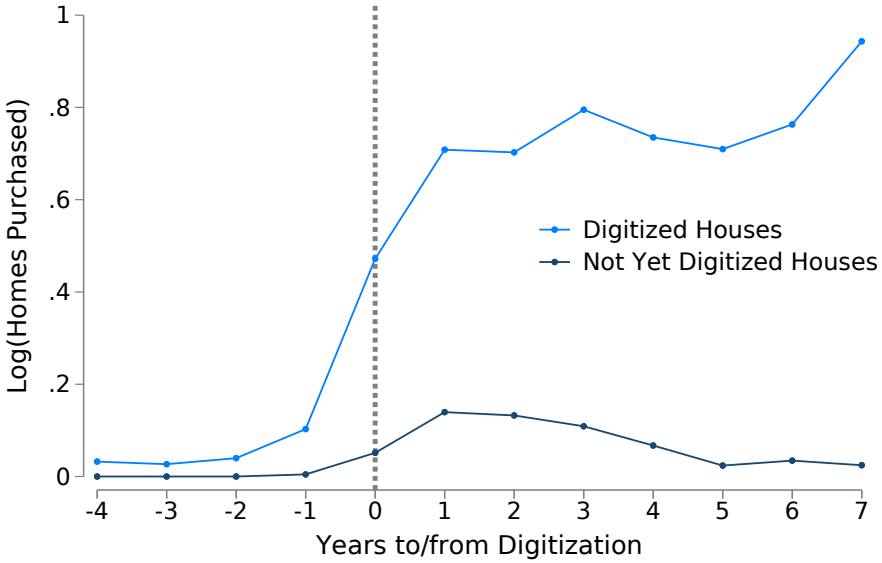
B. ALGORITHMIC INVESTOR SHARE



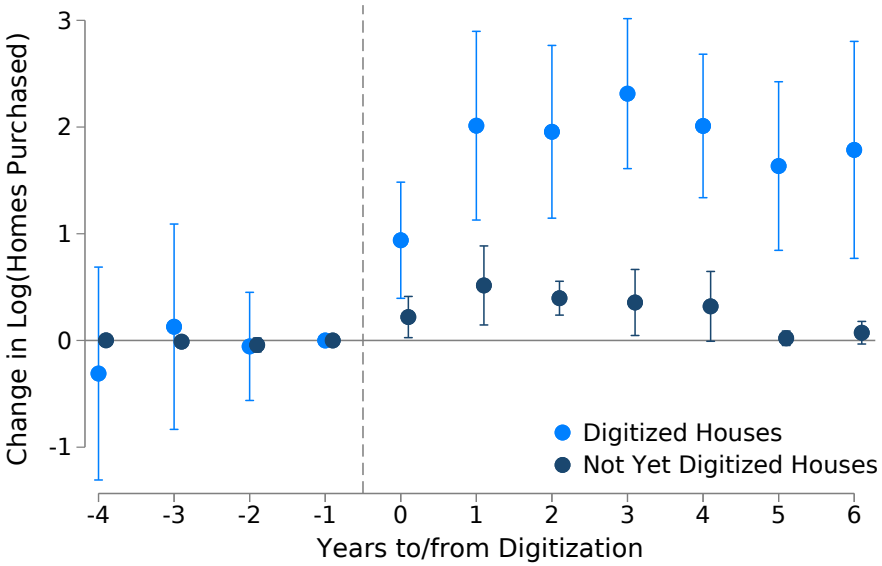
NOTES: These figures plot the coefficients and 95 percent confidence intervals from Sun and Abraham (2021) interaction-weighted event study regressions of county digitization. Panel A shows the natural log of the number of homes purchased by algorithmic investors. Panel B plots the number of algorithmic investors purchases as a share of all transactions. All specifications include state and year fixed effects, standard errors are clustered at the county level. Regressions are weighted by the number of transactions in each county and year. All data come from ATTOM Data, Zillow and county digitization records.

FIGURE 2.4: HOUSES PURCHASED BY ALGORITHMIC INVESTORS, BY HOUSE DIGITIZATION STATUS

A. LOG(HOUSES PURCHASED), BY HOUSE DIGITIZATION, RAW DATA



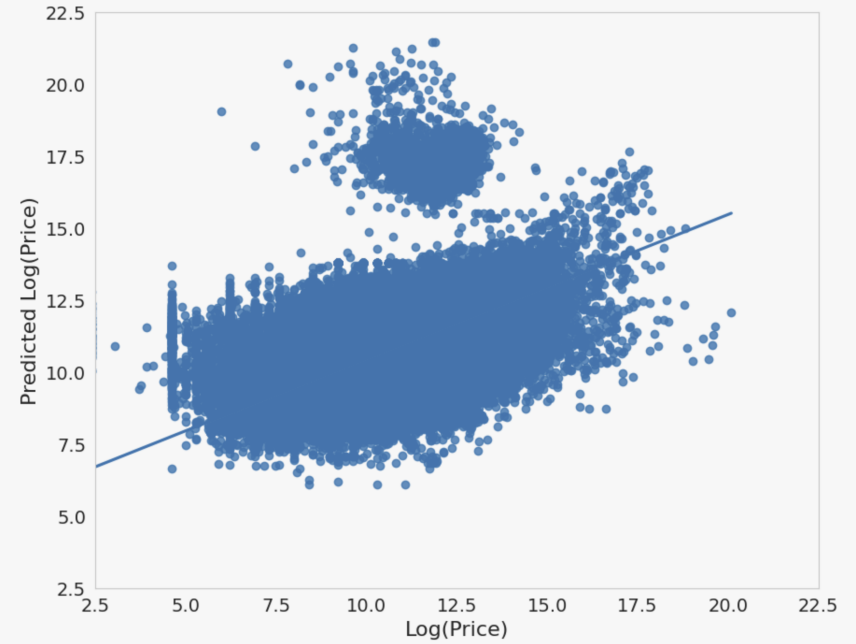
B. LOG(HOUSES PURCHASED), BY HOUSE DIGITIZATION, EVENT STUDY



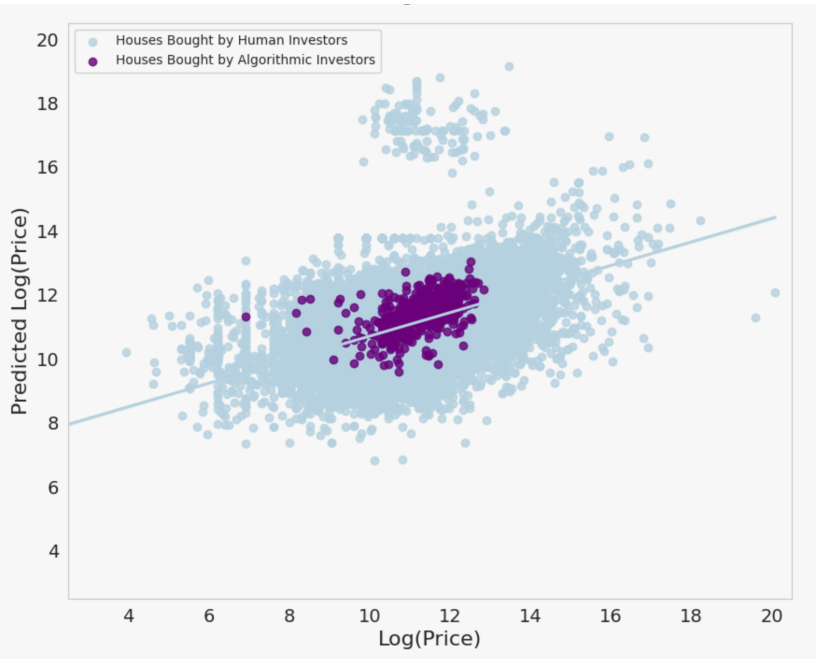
NOTES: These figures show the impact of county digitization on the number of homes purchased by algorithmic firms separately estimated for *digitized houses*, houses that have been digitized, and *non-digitized houses*, houses that have not been digitized and only have paper records. Panel A shows the raw natural log of the number of homes purchased by algorithmic firms and Panel B plots the coefficients and 95 percent confidence intervals from Sun and Abraham (2021) interaction-weighted event study regressions. All specifications include state and year fixed effects, standard errors are clustered at the county level and are weighted by the number of transactions. All data come from ATTOM Data, Zillow and county records.

FIGURE 2.5: MODEL PREDICTED VS. ACTUAL PRICE

A. PREDICTED VS. ACTUAL PRICES, OUT OF SAMPLE

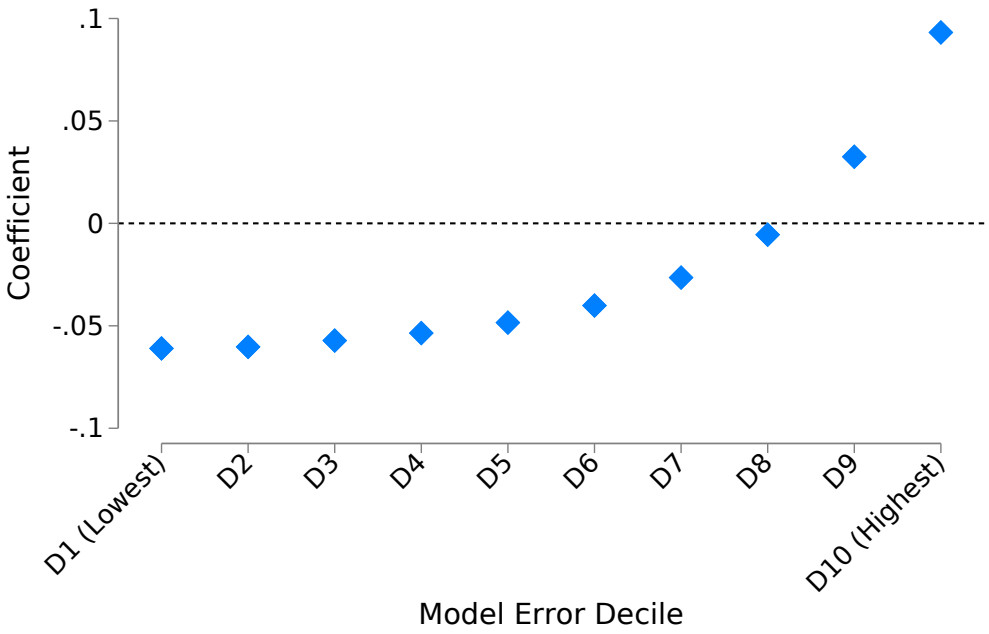


B. PREDICTED VS. ACTUAL PRICES, BY FUTURE INVESTOR PURCHASE



NOTES: Panel A plots the plots the model-predicted natural log of sales price and actual sale price on held out sample of housing transactions. Panel B shows the same results separately for houses that will be purchased in the future by human investors and those that will be purchased by algorithmic investors. All data comes from ATTOM Data and Zillow.

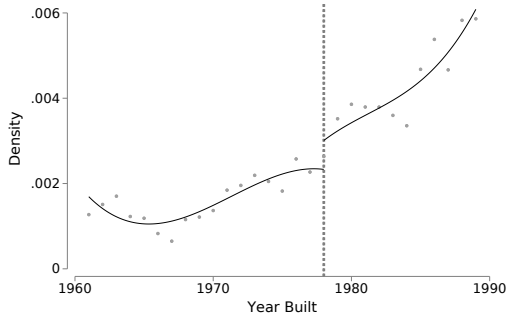
FIGURE 2.6: IMPACT OF DIGITIZATION BY HOUSE PREDICTABILITY



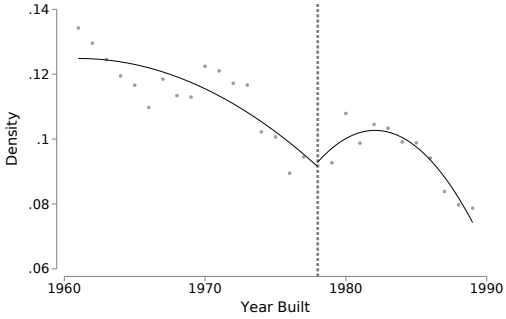
NOTES: These figures plot the impact of house-level digitization on the likelihood of a purchase by a human investor. The model error is calculated as the average difference between the actual and predicted prices for each house. Errors are residualized to account for year-specific fixed effects. Every house is grouped into a decile of model error, with the houses with the lowest mean absolute error in decile 1 and the houses with the largest error in decile 10. All specifications include census block group and year-fixed effects. All data come from ATTOM Data and Zillow.

Figure 2.7: Discontinuities

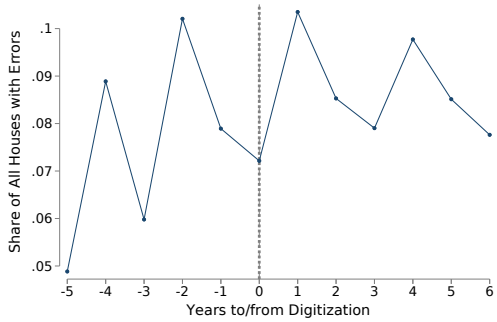
A. Algorithmic Investors



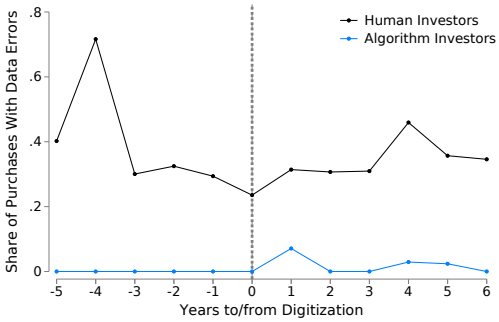
B. Human Investors



C. Share of Houses with Data Errors

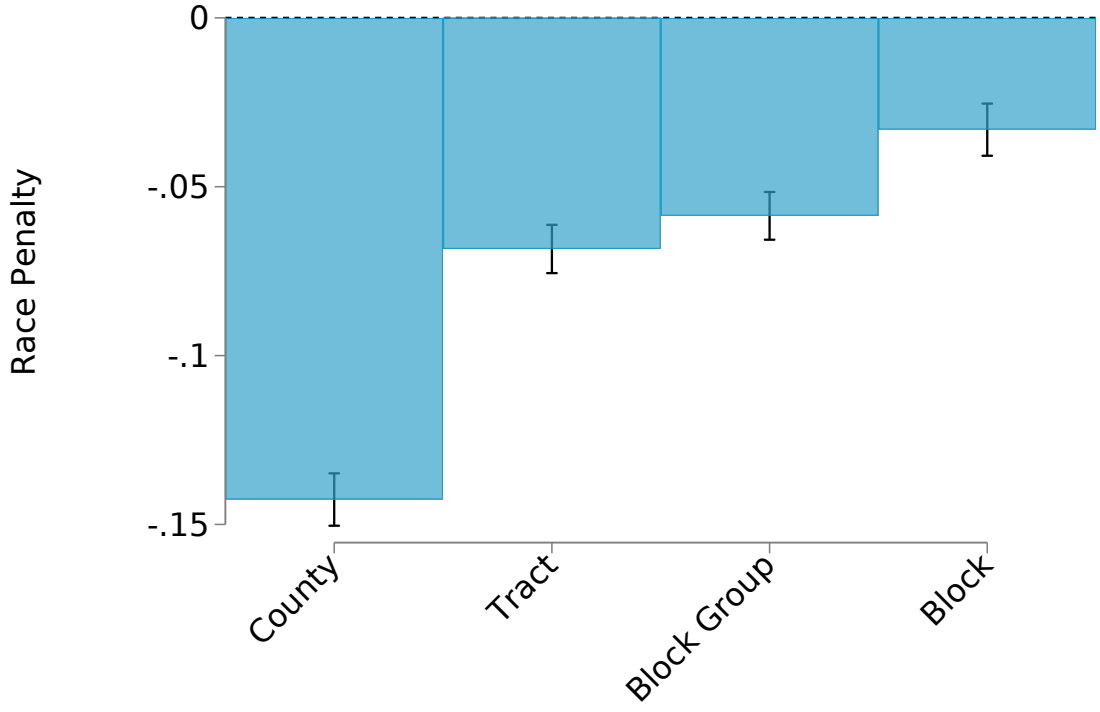


B. Investor Purchases, by Data Errors



Notes: Panel A plots the distribution of houses purchased by algorithmic investors by year of construction. Panel B plots the same for human investors. Panel C shows the share of houses sold every year with data errors. Panel D plots the share of houses purchased by algorithmic and human investors with data errors. All data comes from ATTOM Data and county digitization records.

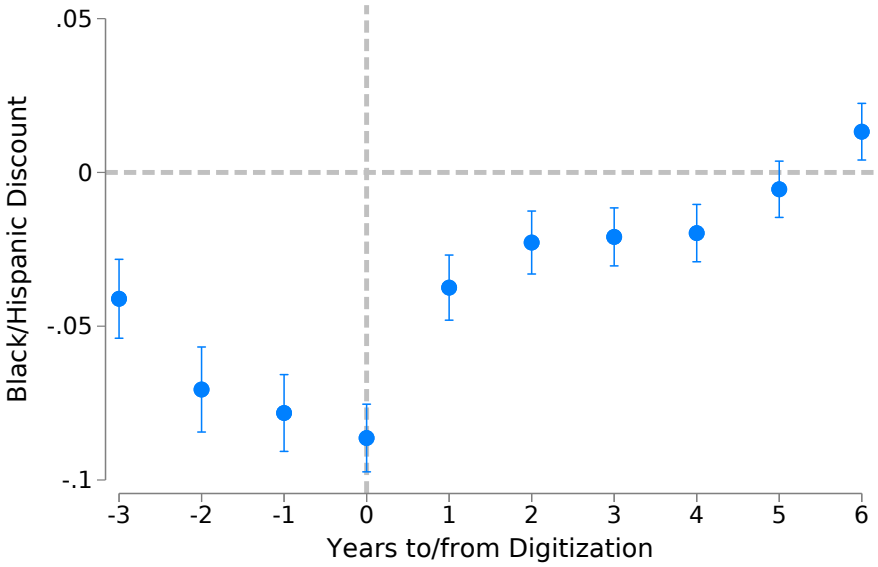
FIGURE 2.8: RACE PENALTY BEFORE DIGITIZATION, BY GEOGRAPHY



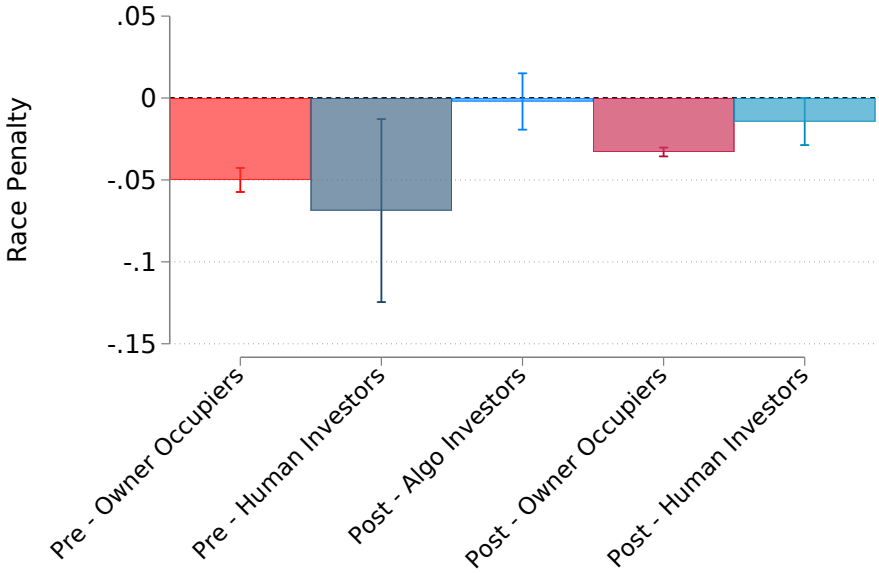
NOTES: This table shows the race penalty or coefficient value that captures the residual difference in sales price between an observably similar house sold by Black or Hispanic homeowners and one sold by a White homeowner. The race penalty is calculated during the time before digitization. The regressions run include geography and year fixed effects along with all available observable characteristics of the house. Standard errors are clustered at the relevant geography. All data comes from ATTOM Data.

FIGURE 2.9: RACE PENALTY, BY TIME TO DIGITIZATION

A. RACE PENALTY, BY TIME TO DIGITIZATION



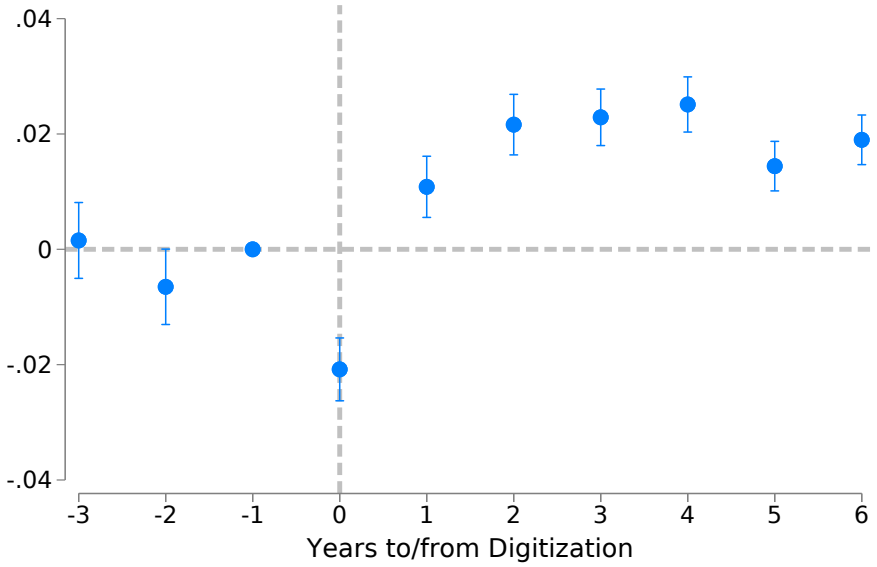
B. RACE PENALTY, BY DIGITIZATION AND BUYER



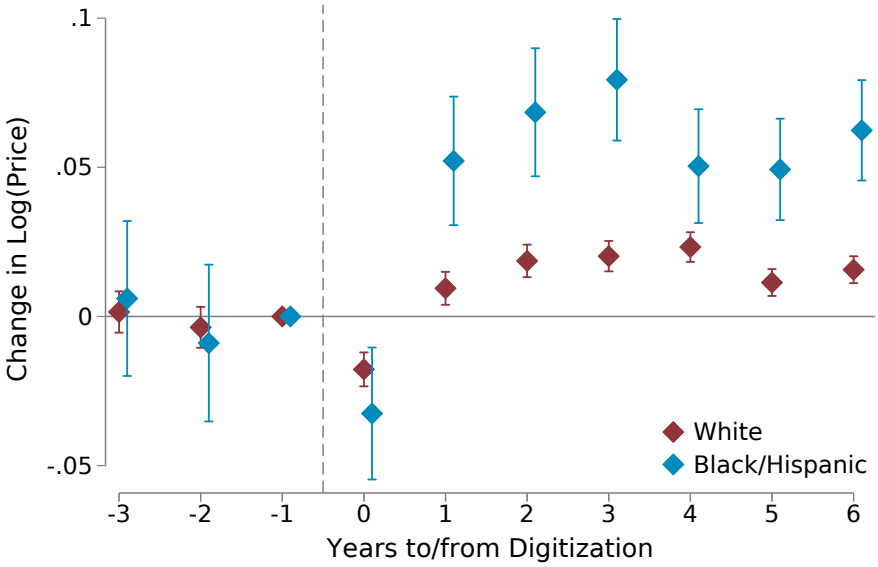
NOTES: Panel A shows the race penalty, or residual difference in sale price between houses sold by White and minority homeowners by time to digitization. All specifications include census block group fixed effects and year-fixed effects and standard errors are clustered at the block group level. Panel B shows the same coefficient plotted pre- and post-digitization for houses purchased by three different types of buyers: owner-occupiers, human investors, and algorithmic investors. All data comes from ATTOM Data and county digitization records.

FIGURE 2.10: DIGITIZATION ON PRICE

A. LOG(PRICE)

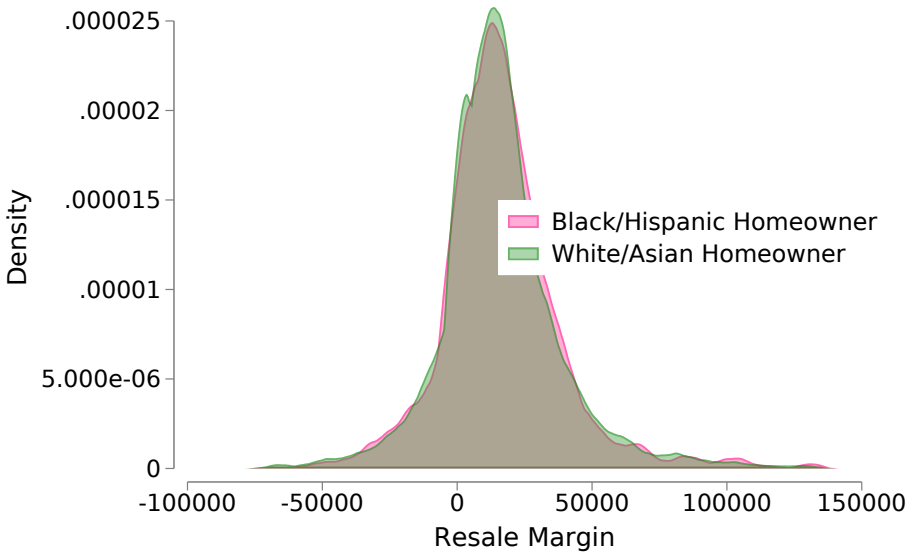


B. LOG(PRICE), BY HOMEOWNER RACE



NOTES: This graph plots the impact of digitization on the natural log of housing transaction prices at the county level in aggregate and separately by White and minority homeowner. All specifications include census block and year fixed effects, standard errors are clustered at the block level. All data comes from ATTOM Data, Zillow and county digitization records.

FIGURE 2.11: RESALE MARGIN, BY HOMEOWNER RACE



NOTES: This graph plots the gross margin or difference between the sale price and the purchase price for houses bought by algorithmic investors according to the race of the homeowner. All data comes from ATTOM Data and county digitization records.

TABLE 2.1: TRANSACTION SUMMARY STATISTICS

	(1) Owner Occupiers	(2) Human Investors	(3) Algo Investors
Sale Price	194,270.04	127,755.99	219,130.88
	(158,431.32)	(145,159.96)	(103,655.74)
Bedrooms	2.12	2.27	2.76
	(3.17)	(3.58)	(1.47)
Bathrooms	2.14	2.09	2.47
	(2.38)	(5.30)	(1.01)
Partial Baths	0.27	0.25	0.43
	(0.48)	(0.48)	(0.50)
Stories	1.25	1.18	1.57
	(0.75)	(0.86)	(0.64)
Additional Buildings	0.07	0.12	0.03
	(0.58)	(1.18)	(0.24)
Garage	0.56	0.48	0.82
	(0.50)	(0.50)	(0.38)
Fireplace	0.59	0.55	0.82
	(0.49)	(0.50)	(0.39)
Basement	0.17	0.13	0.17
	(0.37)	(0.34)	(0.38)
Parking Spaces	0.75	0.58	0.91
	(8.72)	(7.40)	(0.99)
House Age	30.94	36.30	21.31
	(25.89)	(29.26)	(15.53)
Age Since Remodel	24.27	28.82	18.85
	(21.18)	(25.10)	(13.96)
Observations	7223587	975776	111027

mean coefficients; sd in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

NOTES: This table shows the house characteristics of the transactions in our sample. The sample in column 1 includes all houses, including those purchased by owner-occupiers, those buying houses to live in, and investors. Column 2 includes purchases made by human investors, and column 3 includes purchases by investors using algorithms. Houses with missing or zero transaction prices are removed from the sample. All data come from ATTOM Data and ZTRAX.

TABLE 2.2: INVESTORS PURCHASES, BY COUNTY CHARACTERISTICS

Variable	(1) Human Investors	(2) Algorithmic Investors	(3) Difference
County 2010 Population	347,012.69	498,205.72	151,193.02***
	(317,242.69)	(327,782.25)	(0.00)
Total Housing Units	152,197.36	208,074.12	55,876.77***
	(138,687.72)	(141,824.16)	(0.00)
Share Black	27.83	28.58	0.76***
	(16.73)	(14.35)	(0.00)
Share Hispanic	7.77	10.63	2.86***
	(4.20)	(4.66)	(0.00)
Share White	58.71	53.05	-5.66***
	(19.49)	(17.41)	(0.00)
Share Asian	2.98	4.55	1.56***
	(2.29)	(2.93)	(0.00)
Share Persons under 18	24.47	26.39	1.93***
	(2.74)	(2.33)	(0.00)
Median Income	54,298.52	66,305.10	12,006.58***
	(12,705.78)	(12,281.54)	(0.00)
Median Rent	896.24	1,085.91	189.67***
	(191.01)	(182.29)	(0.00)
Share Families in Poverty	11.74	9.31	-2.43***
	(3.82)	(2.83)	(0.00)
Mean Family Size	3.18	3.29	0.10***
	(0.19)	(0.15)	(0.00)
Share Persons under 18	24.47	26.39	1.93***
	(2.74)	(2.33)	(0.00)
Observations	975,776	111,027	1,086,803

NOTES: This table shows socioeconomic and demographic characteristics of counties where algorithmic and human investors purchase houses, weighted by the number of purchases. Data is at the house transaction level. All data comes from the US Decennial Census and the American Community Survey.

Table 2.3: Log(Houses Purchased) by Algorithmic Investors, Difference-in-Difference Estimators

	Point Estimate	Standard Error	Lower Bound 95% Confidence Interval	Upper Bound 95% Confidence Interval
TWFE-OLS	1.130	0.380	0.386	1.874
Borusyak-Jaravel-Spiess	2.451	0.446	1.578	3.325
Callaway-Sant'Anna	1.002	0.021	0.960	1.043
DeChaisemartin-D'Haultfoeuille	2.653	0.325	2.015	3.290
Sun-Abraham	1.988	0.286	1.428	2.549

Notes: This table shows the impact of county data digitization deployment on the log of houses purchased by algorithmic investors. I show results using the robust difference-in-differences estimators introduced in Borusyak, Jaravel and Spiess (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021) along with a traditional two way fixed-effects. Callaway and Sant'Anna (2021) are cannot be weighted, so I present the unweighted estimates. All regressions include county, year fixed effects, and standard errors are clustered at the county level. Regressions are weighted by the number of transactions.

Table 2.4: House Digitization on Algorithmic Investor Purchase

VARIABLES	(1) Algorithmic Investors	(2) Algorithmic Investors	(3) Algorithmic Investors	(4) Algorithmic Investors	(5) Human Investors
House Digitized	0.0023**	0.0021***	0.0009**		
	(0.0011)	(0.0008)	(0.0004)		
County Digitization x House **Not** Digitized				-0.0006	0.0017
				(0.0008)	(0.0023)
County Digitization x House Digitized				0.0098***	-0.0069***
				(0.0008)	(0.0023)
Observations	6,895,957	6,890,606	6,817,554	6,890,606	6,890,606
R-squared	0.0550	0.0598	0.1056	0.0600	0.0716
House Characteristics	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Location FE	Tract	Block Group	Block	Block Group	Block Group
Preperiod DV Mean	.00013	.00013	.00013	.00013	0.1247

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the results of cross-sectional difference-in-difference regressions estimating the impact of house record digitization on the purchase by an algorithmic investor. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the geographic level. All data come from ATTOM Data, ZTRAX and county governments.

TABLE 2.5: RACE PENALTY, WITH HOUSE IMAGES

	(1) Log(Price)	(2) Log(Price)	(3) Log(Price)
Seller Black/Hispanic	-0.0557***	-0.0441***	-0.021***
	(0.0023)	(0.0042)	(0.0051)
Observations	30,130	30,130	29,037
R-squared	0.69	0.71	0.83
House + Lot	Yes	Yes	Yes
Year x Geo	Yes	Yes	Yes
Geographic FE	Tract	Block Group	Block
Adjusted R-squared	.571	.598	.688

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table shows the race penalty–the residual difference in sale price between houses sold by White and minority homeowners. House exteriors are captured using a deep learning model to create vector representations of house images and included in the regressions as controls. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the geographic level. All data come from ATTOM Data, ZTRAX, Zillow and investor websites.

TABLE 2.6: HOUSE DIGITIZATION ON ALGORITHMIC INVESTOR PURCHASE, BY HOMEOWNER RACE

	(1) Algorithm Purchase	(2) Algorithm Purchase	(3) Algorithm Purchase
Seller Minority	-0.0037***	-0.0039***	-0.0049***
	(0.0005)	(0.0005)	(0.0004)
Digitization x Seller White	0.0022**	0.0020**	0.0007*
	(0.0011)	(0.0008)	(0.0004)
Digitization x Seller Minority	0.0044***	0.0042***	0.0043***
	(0.0007)	(0.0006)	(0.0005)
Geography FE	Tract	Block Group	Block
Year FE	Yes	Yes	Yes
Sample	All	All	All
DV Mean	.0018	.0018	.0018
Observations	6895957	6890606	6817554

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

NOTES: This table shows the results of cross-sectional difference-in-difference regressions estimating the impact of house record digitization on the purchase by an algorithmic investor. I separately estimate effects by homeowner race. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the relevant geographic level. All data come from ATTOM Data, ZTRAX and county governments.

Table 2.7: Resale Margin

VARIABLES	(1) Log(Resale Margin)	(2) Log(Resale Margin)	(3) Log(Resale Margin)	(4) Log(Resale Margin)	(5) Log(Resale Margin)	(6) Log(Resale Margin)	(7) Log(Resale Margin)	(8) Log(Resale Margin)
Seller Minority = 1	0.010	0.010	-0.003	0.002	0.095***	0.101***	0.121***	0.066
	(0.009)	(0.010)	(0.018)	(0.017)	(0.019)	(0.021)	(0.033)	(0.041)
Seller Minority x Minority Neighborhood = 1				0.012				0.047
				(0.021)				(0.048)
Observations	6,775	5,630	2,212	5,630	56,459	45,527	23,449	45,527
R-squared	0.459	0.515	0.646	0.515	0.403	0.452	0.474	0.452
FE	Year x Tract	Year x Block Group	Year x Block	Year x Block Group	Year x Tract	Year x Block Group	Year x Block	Year x Block Group
Resale Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Buyers	Algorithms	Algorithms	Algorithms	Algorithms	Humans	Humans	Humans	Humans
DV Mean	0.0646	0.0576	0.0549	0.0576	0.369	0.362	0.307	0.362

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the difference in the natural log of the price the house sells for in the future, or the resale price, and the natural log of the price paid, or the *gross margin*. The *Seller Minority* variable indicates if the house was bought from Black or Hispanic homeowners or White homeowners. *Minority neighborhood* indicates if the house is in a census block group with an above average minority resident share. All specifications includes the resale year and sale year by geography fixed effects, and standard errors are clustered at the geographic level. All data comes from ATTOM Data.

Chapter 3

Hiring as Exploration

Joint with Danielle Li and Peter Bergman

Increasing access to job opportunity for minorities and women is crucial for reducing well-documented race and gender gaps in the economy. While a proliferation of initiatives related to diversity, equity, and inclusion speak to firms' interest in these issues, a persistent doubt remains: how can firms increase diversity without sacrificing quality?

Concerns about "equity–efficiency" tradeoffs in hiring are predicated on the assumption that firms are able to perfectly predict the quality applicants they encounter. In this case, any deviation from the predicted ranking, whether to select more minority or majority group members, would result in a decline in worker quality. In practice, however, an extensive literature has documented that firms, and the recruiters that they employ, are often inaccurate or biased in their predictions (Benson, Li and Shue, 2021; Kline and Walters, 2022). Given that firms appear to be far from perfect in their ability to forecast quality, there may be significant scope for improved evaluation tools to expand opportunities for a broader range of candidates while maintaining or even improving worker quality.

In this paper, we examine the role of algorithms in the hiring process. Resume screening algorithms have become increasingly prevalent in recent years, being used to assess job candidates across various industries and occupations.[1]

Yet while algorithms have been shown to outperform human decision-makers across a range of settings, their use in the hiring process has been controversial, with detractors cautioning that automated approaches may simply codify existing human biases.[2] Amazon, for instance, was widely criticized for using a resume screening algorithm that penalized the presence of the term "women" (for example, "captain of women's crew team") on resumes.[3]

Our paper uses data from a large Fortune 500 firm to study the decision to grant first-round interviews for high-skill positions in consulting, financial analysis, and data science—sectors which offer lucrative jobs with opportunities for career advancement, but which have also been criticized for their lack of diversity. We study the impact of two types of algorithmic approaches: a "supervised learning" model that selects the best candidates as predicted based on its current training data and a "contextual bandit" model that seeks to expand its training data in order to learn about the best candidates over time. Our findings demonstrate that while both algorithmic approaches improve the quality of applicants selected by the firm, they differ in their ability to select diverse candidates. The traditional supervised learning approach leads to a significant reduction in the number of Black and Hispanic workers receiving interviews compared to human hiring practices. In contrast, the contextual bandit approach increases the representation of underrepresented minorities. To our knowledge, this study provides the first empirical evidence that algorithmic design can lead to Pareto

[1] Accurate adoption rates are elusive, but a 2020 survey of human resource executives found that 39% reported using predictive analytics in their hiring processes, a significant increase from just 10% in 2016 (Mercer, 2020). Furthermore, a survey of technology companies indicates that 60% plan to invest in AI-powered recruiting software in 2018, and over 75% of recruiters believe that artificial intelligence will revolutionize hiring practices (Bogen and Rieke, 2018*a*). Throughout this paper, we use the terms "hiring algorithm," "hiring ML," and "resume screening algorithm" interchangeably to refer to algorithms that assist in making initial interview recommendations. It remains rare for algorithms to make final hiring decisions (Raghavan et al., 2019).

[2] For example, see McKinney (2020); Mullainathan and Obermeyer (2019); Russakovsky et al. (2015); Schrittwieser et al. (2019); Yala et al. (2019).

[3] See https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

improvements in both representation and worker quality.

Modern hiring algorithms typically model the relationship between applicant covariates and outcomes in a given training dataset, and then apply this model to predict outcomes for subsequent applicants. By systematically analyzing historical examples, this supervised learning approach can unearth predictive relationships that may be overlooked by human recruiters. Yet because this approach implicitly assumes that past examples extend to future applicants, firms that rely on this approach may favor groups with proven track records, to the detriment of non-traditional applicants. Indeed, because algorithms are most frequently used at the very top of the hiring funnel, this may prevent such applicants from accessing even initial interviews.

We develop and evaluate an alternative algorithm that explicitly values exploration. Our approach begins with the idea that the hiring process can be thought of as a contextual bandit problem: in looking for the best applicants over time, a firm must balance "exploitation" with "exploration" as it seeks to learn the predictive relationship between applicant covariates (the "context") and applicant quality (the "reward"). Whereas the optimal solution to bandit problems is widely known to incorporate some exploration, supervised learning based algorithms engage only in exploitation because they are designed to solve static prediction problems. By contrast, bandit models are designed to solve dynamic prediction problems that involve learning from sequential actions: in the case of hiring, these algorithms value exploration because learning improves future choices.

Our supervised learning model (hereafter, "SL") is based on a logit LASSO that is trained predict an applicant's underlying "hiring potential." We create two models corresponding to two definitions of hiring potential: whether an applicant will receive an offer or be hired (e.g. both receive and accept an offer), if interviewed. Our model is dynamic in the sense that we update its training data throughout our analysis period with the offer and hiring outcomes of the applicants it chooses to interview.[4] This updating allows the SL model to learn about the quality of the applicants it selects, but the model remains myopic in the sense that it does not incorporate the value of this learning into its selection decisions ex-ante.

Our contextual bandit approach implements an Upper Confidence Bound (hereafter, "UCB") algorithm. In contrast to the SL model, which evaluates candidates based on their point estimates of hiring potential, a UCB contextual bandit selects applicants based on the most optimistic assessment of their hiring potential. That is, among applicants with the same predicted hiring potential, the UCB model would prefer the one for whom its estimate is most uncertain. Once candidates are selected, we incorporate their realized offer and hiring outcomes into the training data and update the algorithm for the next period.[5]

In terms of demographics, we show that a traditional SL model would interview substantially fewer Black and Hispanic applicants, relative to the firm's current hiring practices: from 9.4 percent in 4.2 percent. In contrast, implementing a UCB model would more than double the share of interviewed applicants who are Black or Hispanic, from 9.4 percent to 24.3 percent. Both models increase the share of women relative to human recruiters. These

[4]In practice, we can only update the model with data from selected applicants who are actually interviewed (otherwise we would not observe their hiring outcome). See Section 3.4.2 for a more detailed discussion of how this algorithm is updated.

[5]Similar to the SL approach, we are only able to update the UCB model's training data with outcomes for the applicants it selects who are also interviewed in practice. See Section 3.4.2 for a detailed discussion.

results suggest that exploration in the bandit sense—selecting candidates with covariates for which there is more uncertainty—can lead firms to give more opportunities to workers from groups that are under-represented in their training data, even if diversity goals is not a distinct part of the algorithm's mandate.

A key question, however, is what would happen to worker quality. Bandit algorithms may increase demographic representation by exploring, but this exploration could come at the expense of worker quality. To assess this, we must overcome a missing data problem: we only observe hiring outcomes for candidates who were interviewed in reality.[6] We take three complementary approaches, each based on different assumptions, all of which show that algorithms outperform human recruiters in terms of identifying applicants who are more likely to be hired by the firm.

First, we focus on the sample of interviewed candidates for whom we directly observe hiring outcomes. Within this sample, we ask whether applicants preferred by our ML models have a higher likelihood of being hired than applicants preferred by a human recruiter. We find that, for both ML models, applicants with high scores are much more likely to be hired than those with low scores. In contrast, there is almost no relationship between an applicant's propensity to be selected by a human, and their eventual hiring outcome; if anything, this relationship is negative.

Our second approach uses inverse propensity score weighting to recover an estimate of mean hiring likelihood among applicants selected from our full applicant sample. This approach infers hiring outcomes for applicants who are not interviewed using observed outcomes among interviewed applicants with similar covariates and is consistent as long as there is no selection on unobservables. In our setting, this assumption is realistic because human recruiters have access to largely the same resume information we do prior to making an interview decisions and do not interact with applicants beyond the resume screen. We continue to find that ML models improve hiring yield (that is, average hire rates among interviewed applicants): 32 and 27 percent of applicants selected by the SL and UCB models are eventually hired, respectively, compared with only 10 percent among those selected by human recruiters.

Our third approach uses an instrumental variables strategy to address concerns about potential selection on unobservables. In our setting, applicants are randomly assigned to initial resume screeners, who vary in their leniency in granting an interview. We show that applicants selected by stringent screeners (e.g. those subject to a higher bar) have no better outcomes than those selected by more lax screeners: this suggests that humans are not positively screening candidates based on their unobservables. We use this same variation to identify the returns to following ML recommendations on the margin by looking at instrument compliers. We find that marginal candidates with high UCB scores have better hiring outcomes and are also more likely to be Black or Hispanic. Such a finding suggests that following UCB recommendations on the margin would increase both the hiring yield and the demographic diversity of selected interviewees. In contrast, following SL recommendations on the margin would generate similar increases in hiring yield but decrease minority representation.

[6] This is also referred to as a "selective labels" problem. See, for instance, Arnold, Dobbie and Hull (2020); Kleinberg et al. (2018*a*); Lakkaraju et al. (2017).

We also provide some evidence relating hiring yield to other measures of applicant quality. We observe job performance ratings and promotion outcomes for a small subset of workers hired in our sample. Among this selected group, we show that our ML models (trained to maximize hiring likelihood) appear more positively correlated with on the job performance ratings and future promotion outcomes than a model trained to mimic the choices of human recruiters. This provides suggestive evidence that following ML recommendations designed to maximize hiring yield does not come at the expense of on the job performance, relative to following human recommendations.

Finally, we also show that our results are broadly robust to focusing on whether an applicant receives an offer from the firm, rather than whether they are hired (e.g. receive and accept an offer). We repeat much of our initial analysis using models designed to maximize offer likelihood rather than hiring likelihood. Again, we find similar results: relative to human practices, UCB models select a more diverse set of candidates who are also more likely to receive offers. This same is not true for the SL-based offer model which improves offer likelihood but selects fewer minority applicants.

Together, our main findings show that there need not be an equity-efficiency tradeoff when it comes to expanding diversity in the workplace. Specifically, firms' *current* recruiting practices appear to be far from the Pareto frontier, leaving substantial scope for new ML tools to improve both hiring rates and demographic representation. Incorporating exploration in our setting would lead our firm to interview twice as many under-represented minorities while more than doubling its predicted hiring yield. This logic is consistent with a growing number of studies showing that firms may hold persistently inaccurate beliefs about the quality of minority applicants, and may benefit from nudges (algorithmic or otherwise) that generate additional signals of their quality.[7]

At the same time, our SL model leads to similar increases in hiring yield, but at the cost of drastically reducing the number of Black and Hispanic applicants who are interviewed. This divergence in demographic representation between our SL and UCB results demonstrates the importance of algorithmic design for shaping access to labor market opportunities.

In extensions, we consider several alternative screening policies. We show that blinding our algorithms to race and gender variables still generates increases in the share of Black, Hispanic, and female applicants who are selected relative to human hiring. The main difference between our blinded and unblinded UCB models is that the share of selected Asian applicants increases while the share of selected White applicants decreases. We also show that our UCB model performs better on quality when compared to a supervised learning model in which we implement group-specific quotas. Our model further has the advantage of achieving increases in diversity without requiring firms to explicitly specify interview slots by sensitive categories such as race and gender, a practice that often faces legal challenges.

[7]For instance, see Bohren, Imas and Rosenberg (2019); Bohren et al. (2019); Lepage (2020*a*,*b*); Miller (2017).

3.1 Bias in Hiring Practices

3.1.1 Human Hiring

In order to be successful, firms must identify and hire the right workers. In most firms, this task falls to human workers, who screen initial applications for further consideration, conduct interviews, and make final hiring decisions. Because resumes, interviews, and other assessment tools are limited in their ability to reveal an applicant's potential, firms ultimately have to rely on the personal judgment of their recruiters.

A longstanding social sciences literature shows that human evaluators perform their jobs imperfectly. Human decision-makers may be simultaneously cognitively limited in their ability to process data (Benjamin, 2019; Gabaix, 2019; Treisman and Gelade, 1980), over-confident in their assessments (Fischhoff, Slovic and Lichtenstein, 1977; Kausel, Culbertson and Madrid, 2016; Svenson, 1981), and update both too little and too much in response to feedback (Möbius et al., 2022). In addition to these behavioral biases, evaluators may have social preferences for particular applicants. For example, in an ethnographic study, Rivera (2012) documents how recruiters at elite professional services firms favor applicants who share the same hobbies ("She plays squash. Anyone who plays squash I love").

Such behaviors may contribute to already well-documented race and gender gaps in the labor market (Bertrand and Duflo, 2017; Blau and Kahn, 2017; Pager and Shepherd, 2008). For example, research on role congruity theory suggests that managers may find it more difficult to imagine women succeeding in high-level roles because of a mismatch between the qualities stereotypically associated with effective leaders and with women (Eagly and Karau, 2002). Benson, Li and Shue (2021) find, indeed, that managers incorrectly assess women has having lower "potential" within the firm. Similarly, in a large scale correspondence study, Kline and Walters (2022) find evidence that recruiters discriminate against Black applicants across a range of firms and industries. In a study in Eastern Europe, Bartos et al. (2016) show that discrimination in outcomes may be presaged by discrimination in attention: hiring managers pay less attention when evaluating resumes with Roma-sounding names. A variety of papers have explored the mechanisms behind these gaps, with recent work suggesting that some of these differences may be due to managers having incorrect biased beliefs (Bohren, Imas and Rosenberg, 2019).

A variety of studies have considered ways to mitigate these biases, with mixed results. For example, a common suggestion is that women and minorities may benefit by being evaluated by other women and minorities. This solution, however, is often not supported in the data: Bagues and Esteve-Volart (2010), for example, finds that the presence of women on recruiting committees can, in fact, hurt female applicants. Another suggestion is to require decision-makers to undertake anti-bias trainings. While there have been studies showing that de-biasing exercises (perspective taking, counter-stereotyping) can reduce biases in lab settings, there is less evidence about their efficacy in real organizations (Paluck and Green, 2009). Rather, evidence on durable changes in attitudes seem to come from prolonged cross-group exposure (e.g. shared living, schooling, or service) that is difficult for firms to implement as a policy (Bagues and Roth, 2021; Rao, 2019). Finally, affirmative-action approaches that attempt to redress biases against historically marginalized communities by explicitly favoring members of these groups face increasing legal scrutiny (United States

Court of Appeals for the First Circuit, 2020).

Rather than mitigating the biases that evaluators may hold, another strand of research considers the impact of limiting their ability to exercise unconstrained judgment. Hastie and Dawes (2001*b*) survey studies examining the predictive accuracy of human evaluators across a range of settings (graduate school admissions, criminal recidivism, credit risk, and others) and conclude that "expert judgments are rarely impressively accurate and virtually never better than a mechanical judgment rule." Less is known about how constraining human judgment may impact diversity outcomes. In the setting of college admissions, proponents of holistic review have argued that minority groups can benefit from an evaluator's ability to account for assessments of adversity. At the same time, allowing for discretion may introduce opportunities for decisions to be clouded by an evaluator's implicit biases or personal preference (Bertrand, Chugh and Mullainathan, 2005; Prendergast and Topel, 1993).

Rules-based assessments, in essence, suggest that decisions can be improved if humans behaved more like machines. Our paper takes this idea to its conclusion and examines how the growing adoption of algorithms may impact both the quality and equity of firms' hiring practices.

3.1.2 Algorithmic approaches

Firms are increasingly turning toward data-driven tools to improve their hiring practices. These tools range from simple applicant filtering systems that allow recruiters to search for candidates with specific skills or backgrounds, to sophisticated machine learning models designed to predict worker quality. In recent years, a growing number of firms such as HireVue, Pymetrics, and Ideal offer commercial AI-powered tools that can screen resumes and assess candidates' skills. A 2020 industry survey found that 55% of US firms use predictive analytics at some point in their human resource decision-making process, while 41% use algorithms to make predictions about worker fit (Mercer, 2020).[8] Many well-known companies, such as Intel, Johnson and Johnson, Dominos, JP Morgan, United Parcel Service, Mastercard, LinkedIn, Unilever, and Accenture, have openly acknowledged using algorithmic hiring tools for a variety of job roles (Todd, 2019).

The use of algorithms, moreover, is not restricted to large firms who have the resources to buy a customized algorithmic solution. Smaller firms often post their job openings on third party job search platform such as LinkedIn, Indeed, ZipRecruiter and Monster.com, all of which use ML-based tools to decide which applicants to recommend for an open position. As a result, algorithms play a role in screening applicants even for firms that do not themselves employ algorithmic tools.

Hiring algorithms are most commonly used at the "top of the funnel," where they help prioritize applicants for initial interviews. Recruiters at this stage often face the daunting task of sifting through thousands of applications for just a handful of open positions. Firms surveyed about their use of such algorithms frequently express the hope that these tools will enable them to efficiently identify qualified candidates and fill vacancies more quickly (Bogen and Rieke, 2018*b*).

Algorithms may not suffer from some of the key limitations that human recruiters face.

[8] Specifically, pages 38 and 42.

Whereas individual recruiters are likely to base their judgments on their own narrow experience, algorithms are trained on much larger datasets of applicants. For any given applicant, algorithms are able to form predictions using many variables, while human attention is more limited.[9] Algorithms also assess applicants instantaneously, consistently, and without fatigue, in contrast with research showing that human evaluators are inconsistent and suffer from cognitive fatigue (Gabaix, 2019; Hirshleifer et al., 2019).

Consistent with these advantages, existing evidence suggests that algorithms improve the quality of hiring decisions. Hoffman, Kahn and Li (2017) show that the adoption of a job testing technology improves the quality of workers hired in customer service roles and, moreover, that test scores more accurately predict quality than humans. Cowgill (2020) find similar results when considering a resume screening tool for software engineers.

Crucially, however, a growing literature has raised questions about how the growing adoption of algorithms may impact equity and access to job opportunity.[10] A key concern is that algorithms may be trained on data that reflects historical inequities and, in turn, replicate these biases. For example, in a medical setting, Obermeyer et al. (2019) shows that Black patients are less likely to be recommended for medical care than White patients with similar underlying health problems because an algorithm falsely conflates historically lower health spending among Black patients with lower levels of medical need. Anecdotal accounts of algorithmic bias in hiring have also been widely reported in the popular press: an audit of one resume screening model, for instance, found that the two variables it most strongly favored were being named "Jared" and playing high school lacrosse.[11]

Much of this criticism has implicitly focused on algorithms based on "supervised learning." Supervised learning relies on the existence of labeled datasets to train models to predict a given outcome. In the context of hiring, these datasets tend to be based on applicants that a firm has seen and hired in the past. A supervised learning model may then favor applicants who play lacrosse because socioeconomic status or cultural fit has historically been predictive of success in the hiring process. To the best of our knowledge, most commercially available hiring algorithms are based on this type of approach.[12]

In this paper, we highlight an alternative class of algorithms that have thus far not been applied or studied in the context of hiring: contextual bandit algorithms. Whereas supervised learning models focus solely on selecting applicants with high predicted quality, bandit algorithms also seek out candidates in order to learn about their quality. While

[9] Mullainathan and Obermeyer (2019), for instance, provides evidence that the optimal number of variables that predict patient outcomes is greater than the number that doctors can attend to.

[10] For surveys of algorithmic fairness, see Bakalar et al. (2021); Barocas and Selbst (n.d.); Corbett-Davies and Goel (2018); Cowgill and Tucker (2019). For a discussion of broader notions of algorithmic fairness, see Kasy and Abebe (2020); Kleinberg, Mullainathan and Raghavan (2016).

[11] See https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased.

[12] In general, most firms do not provide information on the specifics of their proprietary algorithms. However, several industry sources have indicated that this is true of their own algorithms. Further, most discussions of hiring ML implicitly assume that this is the case. For example, in a survey of firm approaches, Raghavan et al. (2019) discuss many different ways in which firms may implement supervised learning approaches (e.g. what outcomes to train on or what historical data to use), but there is no discussion of any alternative algorithmic approaches that firms may take. We were also unable to find any reports of firms using bandit approaches in our review of various industry surveys, e.g. Bogen and Rieke (2018*b*); Mercer (2020).

Lazear (1998) presents a theoretical model showing that firms can benefit from hiring risky workers, empirical work has largely highlighted uncertainty about a worker's quality as a barrier to hiring (Kuhnen and Oyer, 2016; Sterling and Fernandez, 2018). A small empirical literature, however, has shown that firms can benefit from non-algorithmic policies that push them to adopt more exploratory practices: Miller (2017) shows that temporary affirmative action policies can generate persistent gains in minority representation, while Whatley (1990) documents a similar finding by examining the racial integration of firms following WWI.

We take this idea and ask whether algorithms can implement exploration in a more efficient way. Our approach therefore complements existing work on algorithmic fairness: rather than focusing on the ethical value of diversity, we view diversity as part of exploration which, in turn, is part of optimal learning.

While we apply this approach to resume screening for private sector jobs (described below), we believe these ideas have broader implications for various settings where assessing applicant quality is crucial. These contexts include hiring in academia and the public sector, as well as other selection processes such as credit scoring. In all these cases, the use of supervised learning algorithms implicitly assumes that decision-makers face a static prediction problem.[13] Our paper proposes that these problems could be recast as dynamic learning problems, in which exploration becomes valuable.

3.2 Our Setting

We focus on recruiting for high-skilled, professional services positions, a sector that has seen substantial wage and employment growth in the past two decades (BLS, 2019). At the same time, this sector has attracted criticism for its perceived lack of diversity: female, Black, and Hispanic applicants are substantially under-represented relative to their overall shares of the workforce (Pew Research Center, 2018). This concern is acute enough that companies such as Microsoft, Oracle, Allstate, Dell, JP Morgan Chase, and Citigroup offer scholarships and internship opportunities targeted toward increasing recruiting, retention, and promotion of those from low-income and historically under-represented groups.[14] However, despite these efforts, organizations routinely struggle to expand the demographic diversity of their workforce—and to retain and promote those workers—particularly in technical positions (Athey, Avery and Zemsky, 2000; Castilla, 2008; Jackson, 2020).

We focus on the resume review stage. Openings for professional services roles are often inundated with applications: our firm receives approximately 200 applications for each worker it hires. Interview slots are scarce: because they are conducted by current employees who are diverted from other types of productive work, firms are extremely selective when deciding which of these applicants to interview: our firm rejects 95% of applicants prior to interviewing them.

The process of screening applicants at our firm works as follows. A "hiring manager"

[13] It is important to note that the absence of explicit algorithms does not necessarily imply the absence of potentially biased, algorithm-like thinking. For instance, academics may rely heavily on institutional affiliation when evaluating the quality of a research paper.

[14] For instance, see here for a list of internship opportunities focused on minority applicants. JP Morgan Chase created Launching Leaders and Citigroup offers the HSF/Citigroup Fellows Award.

is in charge of a certain number of job postings (requisitions). The initial interview represents a key barrier for applicants seeking access to employment opportunities. However, recruiters often make these high-stakes decisions quickly, based on limited information, which can lead them to select candidates who ultimately turn out to be unsuitable for the role, while overlooking other candidates who may have excelled. These mistakes not only impact firm productivity but may also perpetuate inequalities by restricting access to economic opportunities. When faced with time pressure, human recruiters may inadvertently rely on heuristics that disadvantage qualified individuals who do not fit traditional models of success (Friedman and Laurison, 2019; Rivera, 2012). In light of these issues, we believe that it is particularly important to understand whether algorithmic tools can be used to improve decisions at the critical initial screening stage.

3.2.1 Data

Our data come from a Fortune 500 company in the United States that hires workers in several job families spanning business and data analytics. All of these positions require a bachelor's degree, with a preference for candidates graduating with a STEM major, a master's degree, and, often, experience with programming in Python, R or SQL. Like other firms in its sector, our data provider faces challenges in identifying and hiring applicants from under-represented groups. We have data on 88,666 job applications from January 2016 to April 2019, as described in Table 3.1. Most applicants in our data are male (68%), Asian (58%), or White (29%). Black and Hispanic candidates comprise 13% of all applications, but under 5% of hires. Women, meanwhile, make up 33% of applicants and 34% of hires.

Applicant covariates

We have information on applicants' educational background, work experience, referral status, basic demographics, as well as the type of position to which they applied. Appendix Table A.1 provides a list of these raw variables, as well as some summary statistics. We have self-reported race (White, Asian, Hispanic, Black, not disclosed and other), gender, veteran status, community college experience, associate, bachelor, PhD, JD or other advanced degree, number of unique degrees, quantitative background (defined having a degree in a science/social science field), business background, internship experience, service sector experience, work history at a Fortune 500 company, and education at elite (Top 50 ranked) US or non-US educational institution. We record the geographic location of education experience at an aggregated level (India, China, Europe). We also track the job family each candidate applied to, the number of applications submitted, and the time between the first and most recent application.

Quality measures

A key challenge our firm faces is being able to hire qualified workers to meet its labor demands; even after rejecting 95% of candidates in deciding whom to interview, 90% of interviews do not result in a hire. These interviews are costly because they divert high-skill current employees from other productive tasks (Kuhn and Yu, 2019). In our paper,

we therefore measure an applicant's quality as their likelihood of either receiving an offer of employment or actually being hired by the firm. By this definition, a high quality applicant is one that meets the firm's own hiring criteria (whatever that may be) or one that meets those criteria and additionally accepts the firm's offer of employment.

Of course, in deciding whom to interview, firms may also care about other objectives: they may look for applicants who have the potential to become superstars—either as individuals, or in their ability to manage and work in teams—or they may avoid applicants who are more likely to become toxic employees (Benson, Li and Shue, 2019; Deming, 2017; Housman and Minor, 2015; Reagans and Zuckerman, 2001; Schumann et al., 2019). Unfortunately, we observe little information on applicants' post-hire performance. For the small set of workers for which we observe this data, we provide noisy evidence that ML models trained to maximize hiring rates are also positively related to performance ratings and promotion rates.

3.3 Empirical Strategy

The goal of our paper is to understand how implementing an exploration-based resume screening algorithm would impact firms' interview outcomes, relative to its existing practices, and relative to traditional supervised learning approaches.

An ideal comparison would involve randomizing screening technologies (human, supervised ML, or exploration ML) through an experiment. In a live implementation, each technology would select which applicants to interview, those applicants would be interviewed, and we would be able to record interview outcomes.

Our analysis, however, relies on archival data. While we observe demographics for all applicants, we only observe quality measures—hiring and offer outcomes—conditional on actually being interviewed. This means that we face a selective labels problem: if an algorithm selects a candidate who is not interviewed in practice, we will not observe that candidate's interview outcome.

In this section, we discuss our framework for addressing this inference challenge. Throughout this section, we will focus on a generic ML-based interview policy. In Section 3.4, we will describe the details of the specific algorithms we implement, but to understand our econometric strategy, it suffices to consider a generic counterfactual.

We begin by formalizing a simple model of a firm's interview decision to fix ideas and notation going forward.

3.3.1 Baseline Framework

We consider a firm that makes interview decisions over time. In each period t, the firm sees a set of job applicants indexed by i, and must choose which candidates to interview $I_{it} \in \{0, 1\}$. The firm would only like to interview candidates that meet its hiring criterion, so a measure of an applicant's quality is her "hiring likelihood": $Y_{it} \in \{0, 1\}$. Y_{it} should be thought of as a potential outcome: would applicant i applying at time t be hired by the firm *if* they were granted an interview? Empirically, Y_{it} as an indicator for whether an applicant receives an offer from the firm or is actually hired (receives and accepts an offer). Regardless

of outcomes, the firm pays a cost, c_t, per interview, which can vary exogenously with time to reflect the number of interview slots or other constraints in a given period.

The firm's payoff for interviewing worker i is given by:

$$Payoff_{it} = \begin{cases} Y_{it} - c_t \text{ if } I_{it} = 1 \\ 0 \text{ if } I_{it} = 0 \end{cases}$$

For each applicant i in period t, the firm also observes a vector of demographic, education, and work history information, denoted by X'_{it}. These variables provide "context" that can inform the expected returns to interviewing a candidate. We write $E[Y_{it}|X'_{it}] = \mu(X'_{it}\theta^*_t)$, where $\mu : \mathbb{R} \to \mathbb{R}$ is a link function and θ^*_t is an unobserved vector describing the true predictive relationship between covariates X'_{it} and hiring potential Y_{it}.[15] We allow X'_{it} to include components that are both observed and unobserved by the econometrician. After each period t, the firm observes the payoffs associated with its chosen actions.

Given this information, we can think of a firm's interview decision for applicant i at time t as given by:

$$I_{it} = \mathbb{I}(s_t(X'_{it}) > c_t) \tag{3.1}$$

Here, $s_t(X'_{it})$ can be thought of as a score measuring the value the firm places on a candidate with covariates X'_{it} at time t. This score is indexed by t to reflect the fact that the value of a given applicant can change over time if the firm's beliefs about their quality change or if the firm's priorities do. The firm's goal is to identify a scoring function $s_t(X'_{it})$ that leads it to identify and interview applicants with $Y_{it} = 1$ as often as possible.

Our model mirrors a standard contextual multi-arm bandit (MAB) problem.[16] Leaving aside the optimal choice of scoring function (which we discuss later in Section 3.4), we can think of *any* interview policy as being described by some scoring function $s_t(\cdot)$ and its associated interview decision I_{it}. In particular, we write s^H_t and I^H_{it} to refer to the (H)uman interview policy that is used by the firm and s^{ML}_t and I^{ML}_{it} to refer to any counterfactual machine-learning (ML) based interview policy. For notational simplicity, we suppress the subscripts for applicant i at time t for the remainder of the paper, unless we are discussing specific regressions or details of algorithm construction.

3.3.2 Addressing sample selection

We are interested in understanding how the quality and demographics of interviewed candidates change under different interview policies. To examine quality, we would like to compare $E[Y|I^H = 1]$ versus $E[Y|I^{ML} = 1]$ for both traditional and exploration based ML approaches. $E[Y|I^H = 1]$ is readily computable because we directly observe hiring potential Y for all workers that are chosen to be interviewed by human recruiters. $E[Y|I^{ML} = 1]$, however, is only partially observable because we only see hiring outcomes for the subset

[15] In practice, when estimating contextual bandit models, most algorithms make functional form assumptions about Most relationship difficult, preventing firms from implementing the ideal decision rule.

[16] In a generic contextual MAB, an agent receives information on "context" before deciding which bandit "arm" to pull in order to receive different "rewards". In our case, the context information is an applicant's resume and demographics X_{it}; the arms are the decision of whether to interview or not $I_{it} \in \{0, 1\}$; and the rewards are the associated payoffs $Y_{it} - c_t$ if I_{it} if interview or 0 if not.

ML-selected applicants who are actually interviewed (e.g. selected by human recruiters): $E[Y|I^{ML} = 1 \cap I^H = 1]$. In our analysis, we address potential biases in assessing counterfactuals associated with sample selection in three complementary ways.

Interviewed sample only

Our first approach compares applicant quality among the subset of applicants who are actually interviewed, for whom we directly observe hiring outcomes. That is, we examine the predictive relationship between algorithm scores s^{ML} and hiring outcomes among the set of applicants with $I^H = 1$. We can also examine the average quality of applicants in the top X% of interviewed applicants, as ranked by our ML scores.

To compare our ML model's preferences to that of human recruiters, we construct a measure of s^H, the implicit "score" that humans assign to applicants by training a model to predict an applicant's likelihood of being selected for an interview $E[I^H|X]$. We are then able to compare the predictive power of $\hat{s}^H$ with that of s^{ML} among the interviewed sample.

Full sample, assuming no selection on unobservables

A concern with our analysis of the interviewed sample is that human recruiters may add value by screening out particularly poor candidates so that they are never observed in the interview sample to begin with. In this case, there may be no correlation between human preferences and hiring potential among those who are interviewed, even though human preferences are highly predictive of quality in the full sample.

Our next approach addresses this by estimating the average quality of *all* ML-selected applicants, $E[Y|I^{ML} = 1]$. We infer hiring likelihoods for ML-selected applicants who were not interviewed using observed hiring outcomes from applicants with similar covariates who were interviewed, assuming no selection on unobservables: $E[Y|I^{ML} = 1, X] = E[Y|I^{ML} = 1, I = 1, X]$.

In our setting, this is a plausible assumption because recruiters have very little additional information relative to what we also observe. Screeners never interact with applicants and make interview decisions on the basis of applicant resumes. Because the hiring software used by our data firm further standardizes this information into a fixed set of variables, they generally do not observe cover letters or even resume formatting. Given this, the types of applicant information that are observable to recruiters but not to the econometrician are predominately related to resume information that we do not code into our feature set. For example, we convert education information into indicator variables for college major designations, institutional ranks, and types of degree. A recruiter, by contrast, will see whether someone attended the University of Wisconsin or the University of Michigan.[17] In addition to worker characteristics, our models also include characteristics of the job search itself to account for factors that influence hiring demand independent of applicant characteristics.

Following Hirano, Imbens and Ridder (2003), we write the inverse propensity weighted

[17] Adding additional granularity in terms of our existing variables into our model does not improve its AUC.

estimate of ML-selected workers' hiring likelihood as:

$$
\begin{aligned}
E[Y|I^{ML}=1] &= \sum_X p(X|I^{ML}=1)E[Y|I^{ML}=1,X] \\
&= \sum_X \frac{p(I^{ML}=1|X)p(X)}{p(I^{ML}=1)}E[Y|I^{ML}=1,X] \\
&= \frac{1}{p(I^{ML}=1)}\sum_X p(I^{ML}=1|X)p(X)E[Y|I^{ML}=1,X]\frac{p(X|I=1)p(I=1)}{p(I=1|X)p(X)} \\
&= \frac{p(I=1)}{p(I^{ML}=1)}\sum_X E[Y|I=1,X]\frac{p(I^{ML}=1|X)p(X|I=1)}{p(I=1|X)} \\
&\quad \text{(Assuming selection on observables)} \\
&= \frac{p(I=1)}{p(I^{ML}=1)}E\left[\frac{p(I^{ML}=1|X)}{p(I=1|X)}Y|I=1\right] \qquad (3.2)
\end{aligned}
$$

Equation (3.2) says that we can recover the mean quality of ML-selected applicants by reweighting outcomes among the human-selected interview sample, using the ratio of ML and human-interview propensity scores. The ML decision rule is a deterministic function of covariates X, meaning that the term $p(I^{ML}=1|X)$ is an indicator function equal to one if the ML rule would interview the applicant, and zero if not. The term $p(I^H=1|X)$ is just the human selection propensity which we estimate as $\hat{s}^H$, described previously. Finally, because we always select the same number of applicants as are actually interviewed in practice, the term $\frac{p(I=1)}{p(I^{ML}=1)}$ is equal to one by construction.

Selection on unobservables and IV analysis

Our next approach addresses concerns about selection on unobservables. We are particularly concerned about the case in which there is positive selection by human recruiters, or human interviewers screen out applicants with unobservably bad covariates. In this case, our previous approach would overstate the quality of ML-selected applicants who did not receive an interview.

To address this concern, we take advantage of random assignment of applicants to initial resume screeners, following the methodology pioneered by Kling (2006). Resume screeners in our data vary in their propensity to pass applicants to the interview round: an applicant may receive an interview if she is assigned to a generous screener and that same applicant may not if she is assigned to a stringent one. For each applicant, we form the jackknife mean pass rate of their assigned screener and use this as an instrument, Z, for whether the applicant is interviewed. In Section 3.5.2, we provide evidence for the validity of this instrument. Assuming it is valid, this instrument aids our analysis in two ways.

First, it allows us to directly test for the presence of selection on unobservables. If humans are, on average, positively selecting candidates, then it should be the case that applicants selected by more stringent reviewers—e.g. those who are subjected to a higher bar—should be more likely to be hired conditional on being interviewed than those selected by more lenient reviewers. That is, if there is a positive relationship between human selection

propensities and hiring outcomes, then going further down the distribution by selecting more candidates should decrease average quality. We will show that this is not the case.

Second, our instrument allows us to consider an alternative approach for valuing the performance of ML models relative to human decisions: instead of comparing hiring outcomes across the full sample (which requires that we assume no selection on unobservables), we show that firms can improve hiring yield by relying on algorithmic recommendations in cases where human screeners are on the margin of granting an interview.

To demonstrate this, consider the following counterfactual interview policy, given our recruiter leniency instrument Z:

$$\tilde{I} = \begin{cases} I^{Z=1} \text{ if } s^{ML} \geq \tau, \\ I^{Z=0} \text{ if } s^{ML} < \tau. \end{cases}$$

The policy $\tilde{I}$ takes the firm's existing interview policy, I, and modifies it at the margin. The new policy $\tilde{I}$ favors applicants with high ML scores by asking the firm to make interview decisions I as if these applicants were randomly assigned to a generous initial screener ($Z = 1$).[18] That is, $I^{Z=1}$ refers to the counterfactual interview outcome that would be obtained, if an applicant were evaluated by a lenient screener. Similarly, $\tilde{I}$ penalizes applicants with low ML scores by making interview decisions for them as though they were assigned to a stringent screener ($Z = 0$).

By construction, the interview policy $\tilde{I}$ differs from the status quo policy I only in its treatment of instrument compliers. Compliers with high ML scores will be selected under $\tilde{I}$ because they are always treated as if they are assigned to lenient recruiters. Conversely, compliers with low ML scores are always rejected because they are treated as if they are assigned to strict reviewers. As such, the returns to following ML recommendations on the margin is determined by whether compliers with high ML scores have greater hiring potential than compliers with low ML scores, $E[Y|I^{Z=1} > I^{Z=0}, s^{ML} \geq \tau]$ vs. $E[Y|I^{Z=1} > I^{Z=0}, s^{ML} < \tau]$.

To compute the hiring potential of compliers, we estimate regressions following Benson, Li and Shue (2019) and Abadie (2003):

$$Y_{it} \times I_{it} = \alpha_0 + \alpha_1 I_{it} + X'_{it}\alpha + \varepsilon_{it} \text{ if } s^{ML}(X'_{it}) \geq \tau \quad (3.3)$$

$$Y_{it} \times I_{it} = \beta_0 + \beta_1 I_{it} + X'_{it}\beta + \varepsilon_{it} \text{ if } s^{ML}(X'_{it}) < \tau \quad (3.4)$$

In Equation (3.3), $Y_{it} \times I_{it}$ is equal to applicant it's hiring outcome if she is interviewed or to zero if she is not. This regression is structured so that the OLS coefficient $\hat{\alpha}_1^{OLS}$ estimates the average hiring potential among all interviewed applicants with high ML scores. The IV estimate $\hat{\alpha}_1^{IV}$, in contrast, is an estimate of hiring potential among high ML-score compliers: $E[Y_{it}|I_{it}^{Z_{it}=1} > I_{it}^{Z_{it}=0}, s^{ML} \geq \tau]$. Similarly, $\hat{\beta}_1^{IV}$ in Equation (3.4) is the analogous estimate for low ML-score compliers: $E[Y_{it}|I_{it}^{Z_{it}=1} > I_{it}^{Z_{it}=0}, s^{ML} < \tau]$. This logic is analogous to the idea that IV estimates identify a LATE amongst compliers.[19]

[18] For simplicity in exposition, we let Z be a binary instrument in this example (whether an applicant is assigned to an above or below median stringency screener) although in practice we will use a continuous variable.

[19] In standard potential outcomes notation, the LATE effect is $E[Y^1 - Y^0|I^{Z=1} > I^{Z=0}]$. In our case,

3.4 Algorithm Design

Having discussed our general empirical strategy, we now provide an overview of the specific algorithms we consider.

3.4.1 Preliminaries

We begin by clarifying some issues relevant for all models we consider.

We divide our data into two periods, the first consisting of the 48,719 applicants that arrive before 2018 (2,617 of whom receive an interview), and the second consisting of the 39,947 applications that arrive in 2018-2019 (2,275 of whom are interviewed). We think of the 2016-2017 period as our "training" data and the 2018-2019 period as our "analysis" period. This approach to defining a training dataset (rather than taking a random sample of our entire data) most closely approximates a real world setting in which firms would likely use historical data to train a model that is then applied prospectively. Once we have trained our baseline model, we will continue to update our models using information on candidates they select during the 2018-2019 sample. Specifically, we divide our analysis sample into "rounds" of 100 applicants. After each round, we take the applicants the model has selected and update its training data. We then retrain the model and use its updated predictions to make selection decisions in the next round. This will be described in more detail below.

Our goal is to predict applicants' hiring potential, Y, as defined in Section 3.3.1. Because hiring potential is only directly observed for applicants who are interviewed, we train our models using data from interviewed applicants only. We note that not all screening algorithms are trained in this way. One common approach is to predict the human recruiter's interview decision, I, rather than the applicant's hiring potential Y. In this case, vendors are able to train their data on all applicants. Another common approach is to focus on predicting hiring likelihood Y, but to set $Y = 0$ for all applicants who are not interviewed. This essentially assumes that candidates who were not interviewed had low hiring potential. We choose not to follow either of these approaches because they conflate a recruiter's decision I with the quality of that decision Y. Rambachan and Roth (2019), in addition, show that such approaches tend to be more biased against racial minorities.

Second, our primary measures of quality—hire and offer outcomes—are based on the discretion of managers and potentially subject to various types of evaluation and mentoring biases (Castilla, 2011; Quadlin, 2018). Indeed, measures of on the job performance such as performance ratings or promotion rates would also be subject to this concern. Without a truly "objective" measure of quality, we interpret our results as asking whether ML tools can improve firm decisions, as measured by its own revealed preference metrics (whether it chooses to hire, promote, or rate someone highly). Most available hiring algorithms focus on finding applicants firms would like to hire and often lack comprehensive on-the-job performance data, so our setup closely mirrors algorithms used in practice.

Third, our models may generate inaccurate predictions if the relationship between covari-

we are only interested in the average potential outcome of compliers: $E[Y^1|I^{Z=1} > I^{Z=0}]$. Here, Y^1 is equivalent to a worker's hiring outcome if she is interviewed—this is what we have been calling quality, H. For a formal proof, see Benson, Li and Shue (2019).

ates X and hiring likelihood Y differs between the full applicant sample and the interviewed subset. While there are a growing set of advanced ML tools that seek to correct for training-sample selection,[20] testing these approaches is outside of the scope of this paper and we are not aware of any commercially available algorithms that employ sample selection correction. In Section 3.5.2, we use an IV approach to provide evidence that selection on unobservables does not appear to be a large concern in our data.

In general, we emphasize that the ML models we build should not be thought of as "optimal" in either their design or performance, but as an example of what could be feasibly achieved by most firms that able to organize their administrative records into a modest training dataset, with a standard set of resume-level input features, using a technically accessible ML toolkit.

Supervised Learning ("SL")

Our first model uses a standard supervised learning approach to predict an applicant's likelihood of being hired, conditional on being interviewed. We begin with an initial training dataset, D_0 and use it to form an estimate of applicant quality $\hat{E}[Y_{it}|X'_{it}; D_0]$. Specifically, we use a L1-regularized logistic regression (LASSO), fitted using three-fold cross validation, trained on job applicants from the first half of our sample, 2016-2017. Appendix Figure A.1 plots the receiver operating characteristic (ROC) curve and its associated AUC, or area under the curve. This model has an AUC of 0.64, meaning that it will rank an interviewed applicant who is hired ahead of an interviewed but not hired applicant 64 percent of the time.[21]

Having trained this initial model on 2016-17 data, we use it to make interview decisions for future applicants. If the firm always uses the same model—that is, if it never again updates its training data—we say that the firm follows a "static" SL model. This approach is quite common in practice.

In our paper, however, we estimate a dynamic SL model that updates the firm's training data with the outcomes of applicants it selects later on. Specifically, we divide our analysis sample (2018-2019) into "rounds" of 100 applicants. After each round, we take the applicants the model has selected and update its training data. We then retrain the model and use its updated predictions to make selection decisions in the next round. At any given point t, the SL model's interview policy is as follows, based on Equation (3.1) of our conceptual framework:

$$I_{it}^{SL} = \mathbb{I}(s_t^{SL}(X'_{it}) > c_t), \text{ where } s_t^{SL}(X'_{it}) = \hat{E}[Y_{it}|X'_{it}; D_t^{SL}]. \tag{3.5}$$

Here, D_t^{SL} is the training data available to the algorithm at time t.

It is important to emphasize that we can only update the model's training data with *observed* outcomes for the set of applicants selected in the previous period: that is, $D_{t+1}^{USL} = D_t^{USL} \cup (I_t^{USL} \cap I_t)$. Because we cannot observe hiring outcomes for applicants who are not

[20] See, for example, Dimakopoulou et al. (2018*a*), Dimakopoulou et al. (2018*b*) which discuss doubly robust estimators to remove sample selection and Si et al. (2020).

[21] The AUC is a standard measure of predictive performance that quantifies the tradeoff between a model's true positive rate and its false positive rate. Formally, the AUC is defined as $\Pr(s(X'_{it}) > s(X'_{jt})|Y_{it} = 1, H_{jt} = 0)$. We also plot the confusion matrix in Appendix Figure A.2, which further breaks down the model's classification performance.

interviewed in practice, we can only update our data with outcomes for applicants selected by both the model and by actual human recruiters. This may impact the degree to which the SL model can learn about the quality of the applicants it selects, relative to a world in which hiring potential is fully observed for all applicants and we discuss this in more detail shortly, in Section 3.4.2.

Upper Confidence Bound ("UCB")

While there is in general no generic optimal strategy for the contextual bandit model described in Section 3.3.1, it is widely known that exploitation-only approaches—such as the SL model described above—are inefficient solutions because they do not factor the ex-post value of learning into their ex-ante selection decisions (Bastani, Bayati and Khosravi, 2019; Dimakopoulou et al., 2018*b*). An emerging literature in computer science has therefore focused on developing a range of computationally tractable algorithms that incorporate exploration.[22]

The particular exploration-based implementation we use is an Upper Confidence Bound Generalized Linear Model (UCB-GLM) described in Li, Lu and Zhou (2017). We choose this approach because it best fits our setting. UCB-GLM works well when the relationship between "context" variables (covariates X_{it}) and "reward" (hiring potential, Y_{it}) follows a generalized linear functional form ($E[Y_{it}|X'_{it}] = \mu(X'\theta^*_t)$): we measure Y_{it} as a binary hiring outcome and estimate $E[Y_{it}|X'_{it}]$ using a logistic regression. Under these circumstances, Li, Lu and Zhou (2017) provides the algorithm implementation we follow and shows that it is asymptotically regret-minimizing.[23]

Specifically, our UCB algorithm scores applicant i in period t as follows:

$$I^{UCB}_{it} = \mathbb{I}(s^{UCB}_t(X'_{it}) > c_t), \text{ where } s^{UCB}_t(X'_{it}) = \hat{E}[Y_{it}|X'_{it}; D^{UCB}_t] + \alpha B(X'_{it}; D^{UCB}_t). \quad (3.6)$$

In Equation (3.6), the scoring function $s^{UCB}_t(X'_{it})$ is a combination of the algorithm's expectations of an applicant's quality based on its training data and an "exploration bonus" given by:

$$B(X'_{it}; D^{UCB}_t) = \sqrt{(X_{it} - \bar{X}_t)' V_t^{-1} (X_{it} - \bar{X}_t)}, \text{ where } V_t = \sum_{j \in D^{UCB}_t} (X_{jt} - \bar{X}_t)(X_{jt} - \bar{X}_t)'. \quad (3.7)$$

Intuitively, Equation (3.6) breaks down the value of an action into an exploitation component and an exploration component. In any given period, a strategy that purely focuses

[22]The best choice of algorithm for a given situation will depend on the number of possible actions and contexts, as well as on assumptions regarding the parametric form relating context to reward. For example, recently proposed contextual bandit algorithms include UCB (Auer (2002)), Thompson Sampling (Agrawal and Goyal (2013)), and LinUCB (Li et al. (2010)). In addition, see Agrawal and Goyal (2013), and Bastani and Bayati (2019). Furthermore, the existing literature has provided regret bounds—e.g., the general bounds of Russo and Roy (2015), as well as the bounds of Rigollet and Zeevi (2010) and Slivkins (2014) in the case of non-parametric function of arm rewards—and has demonstrated several successful applications areas of application—e.g., news article recommendations (Li et al. (2010)) or mobile health (Lei, Tewari and Murphy (2017)). For more general scenarios with partially observed feedback, see Rejwan and Mansour (2019) and Bechavod et al. (2020).

[23]See Equation 6 and Theorem 2 of their paper.

on exploitation would choose to interview a candidate on the basis of her expected hiring potential: this is encapsulated in the first term, $\hat{E}[Y_{it}|X'_{it}; D_t^{UCB}]$. Indeed, this is essentially the scoring function for the SL model, described in Equation (3.5). Meanwhile, a strategy that purely focuses on exploration would choose to interview a candidate on the basis of the distinctiveness of her covariates: this is encapsulated in the second term, $B(X'_{it}; D_t^{UCB})$, which shows that applicants receive higher bonuses if their covariates deviate from the mean in the population $(X_{it} - \bar{X}_t)$, especially for variables X'_{it} that generally have little variance in the training data (e.g. weighted by the precision matrix V_t^{-1}). To balance exploitation and exploration, Equation (3.6) combines these two terms. As a result, candidates are judged on their mean expected quality *plus* their distinctiveness from the existing training data. Taken together, Li, Lu and Zhou (2017) shows that this provides an upper bound on the confidence interval associated with an applicant's true quality, given the training data D_t^{UCB}, hence the term UCB. In essence, UCB approaches are based on the principle of "optimism in the face of uncertainty," favoring candidates with the highest statistical upside potential.[24]

At time $t = 0$ of our analysis sample, our UCB and SL models share the same predicted quality estimate, which is based on the baseline model trained on the 2016-2017 sample. As with the SL model, we update the UCB model's training data with the outcomes of applicants it has selected during the 2018-2019 analysis period. Based on these new training data, the UCB algorithm updates both its beliefs about hiring potential and the bonuses it assigns. As was the case with the SL model, we can only add applicants who are selected by the model and also interviewed in practice. We now turn to the implications of this limitation.

Model comparisons

A large theoretical literature shows that exploration-based model such as UCB will outperform exploitation-only based approaches such as SL in the long run (Dimakopoulou et al., 2018*b*). Li, Lu and Zhou (2017) proves that the specific model we adopt, UCB-GLM will asymptotically minimize regret via more efficient learning: that is, it will select applicants with greater hiring potential. Yet while a UCB based approach is expected to out-perform SL models in the long run, the quality differences we will observe in practice are ambiguous and capture both the long term benefits of learning and the short term costs of exploration. This tradeoff will also depend on the specifics of our empirical setting. In particular, if quality is not evolving and there is relatively rich initial training data, SL models may perform as well as if not better than UCB models because the value of exploration will be limited. If, however, the training data were sparse or if the predictive relation between context and rewards evolves over time, then the value of exploration is likely to be greater.

In terms of diversity, our UCB algorithm favors candidates with distinctive covariates because this helps the algorithm learn more about the relationship between applicant covariates and hiring outcomes. This suggests that a UCB model would, at least in the short run, select more applicants from demographic groups that are under-represented in its training

[24]The basic UCB approach for non-contextual bandits was introduced by Lai and Robbins (1985) and, since then, various versions of this approach have been developed for different types of contextual bandit settings, and shown to be regret minimizing.

data, relative to an SL model.[25] Over time, however, exploration bonuses will decline as the model receives more information about applicants of all types. As a result, long run differences in selection patterns between SL and UCB models will be driven by differences in their beliefs about applicant quality. Gains in diversity driven by exploration bonuses will not be sustainable if minority applicants are actually weaker.

Our main results will come from a 16 month period from January 2018 to April 2019. Because most organizations interested in adopting screening algorithms are unlikely to operate over a very long time horizon, it becomes important to empirically examine how exploration-based algorithms behave over medium-run time scales. To assess the performance of the algorithms with changes in applicant quality, we supplement our main results with a series of simulations.

3.4.2 Feasible versus Live Model Implementation

In a live implementation, each algorithm would select which applicants to interview, and the interview outcomes for these applicants would be recorded. Our retrospective analysis on quality is limited by the fact that we only observe interview outcomes for applicants who were actually interviewed (as chosen by the firm's human screeners) and, as such, we are only able to update our USL and UCB models with outcomes for candidates in the intersection of human and algorithmic decision making. Here, we discuss how the actual interpretation of our models—which we term "feasible" USL or UCB—may differ from a live implementation.

For concreteness, suppose that the UCB model wants to select 50 Black applicants with humanities degrees in order to explore the applicant space. But, in practice, only 5 such applicants are actually interviewed. In our feasible implementation, we would only be able to update the UCB's training data with the outcomes of these five applicants, whereas in a live implementation, we would be able to update with outcomes for all 50 UCB-selected candidates.

If there is no selection on unobservables on the part of the human recruiters, the feasible UCB model's estimate of the quality of Black humanities majors would be the same as the live UCB's estimate but, because it observes five rather than 50 instances, its estimates would be considerably less precise. This uncertainty would show up in the next period via the exploration bonus term of Equation (3.6): even though it has the same beliefs about quality, the feasible UCB would likely select more Black humanities majors in the next period relative to a live UCB because the standard error around its estimate would be greater. In this way, selection on observables should be thought of as slowing down the process of learning. In the limit, the feasible and live models should converge to the same beliefs and actions because, with a large enough sample, there would be little uncertainty driving exploration bonuses.[26]

[25] We allow bonuses to be calculated over all covariates we observe. If demographic minorities are more homogenous along other dimensions such as educational background and work history, then this would decrease the correlation between minority status and exploration bonuses. For example, Asian applicants make up the majority of our applicant sample and so would receive low exploration bonuses on the basis of race alone; however, they are more likely to have non-traditional work histories or have gone to smaller international colleges, factors that make them more distinctive to the UCB model.

[26] Formally, the distinction between the feasible and live versions of our ML models is related to regression in which outcomes are missing at random conditional on unobservables. Under the assumption of no selection

This analysis changes if there is human interviewers select based on unobservables. A particularly concerning version occurs if this selection is positive. In this case, the 5 Black humanities majors that are actually interviewed will tend to be higher quality than the 45 Black humanities majors who were not interviewed. A feasible UCB model will then be too optimistic about the quality of this population relative to a live UCB model that learns about the quality of all 50 such applicants. In the next period, the feasible UCB model will select more Black humanities majors both because uncertainty for these applicants remains higher and because selection on unobservables induces upwardly biased beliefs. This latter bias can lead our approach to select too many applicants from groups whose weaker members are screened out of the model's training data by human recruiters.

In Section 3.5.2 we provide IV-based evidence that human recruiters do not appear to be selecting on unobservables. In addition, Section 4.5.2 shows simulation results in which we update outcomes for all ML-selected candidates. This allows us to explore the learning behavior of our USL and UCB models in settings that more closely approximate a live implementation.

3.5 Main Results

3.5.1 UCB and SL versus Human Recruiters: Diversity of Selected Applicants

We begin by assessing the impact of each policy on the diversity of candidates selected for an interview in our analysis sample. This is done by comparing $E[X|I = 1]$, $E[X|I^{SL} = 1]$, and $E[X|I^{UCB} = 1]$, for various demographic measures X, where we choose to interview the same number of people as the actual recruiters in a given year-month. We observe demographic covariates such as race and gender for all applicants, regardless of their interview status.

We focus on the racial composition of selected applicants. Panel A of Figure 3.1 shows that, at baseline, 54% of applicants in our analysis sample are Asian, 25% are White, 8% are Black, and 4% are Hispanic. Panel B shows that human recruiters select a similar proportion of Asian and Hispanic applicants (57% and 4%, respectively), but relatively more White and fewer Black applicants (34% and 5%, respectively). In Panel C, we show that the SL model reduces the share of Black and Hispanic applicants from 10% to under 5%, White representation increases more modestly from 34% to 42%, and Asian representation stays largely constant. In contrast, Panel D shows that the UCB model increases the Black share of selected applicants from 5% to 15%, and the Hispanic share from 4% to 9%. The White share stays constant, while the Asian share falls from 57% to 44%.

Appendix Figure A.5 plots the same set of results for gender. Panel A shows that 65% of interviewed applicants are men and 35% are women; this is largely similar to the gender composition of the overall applicant pool. Unlike the case of race, both our ML models are

on unobservables, common support, and well-specification of the regression function (in our case, the logit), the feasible and live versions of our models should both be consistent estimators of the underlying parameter θ^* linking covariates with hiring outcomes: $E[Y_{it}|X'_{it}] = \mu(X'_{it}\theta^*_t)$ (Robins, Rotnitzky and Zhao, 1995; Wang, Rotnitzky and Lin, 2010). In a finite sample, of course, the point estimates of the feasible and live models may differ.

aligned in selecting more women than human recruiters, increasing their representation to 41% (SL) or 44% (UCB).

Next, we explore why our UCB model selects more Black and Hispanic applicants. Panel A of Appendix Figure A.6 shows that Black and Hispanic applicants receive slightly larger exploration bonuses on average. This reflects both direct differences in population size by race, as well as indirect differences arising from the correlation between race and other variables that also factor into bonus calculations. Appendix Figure A.7 plots the proportion of the total variation in exploration bonuses that can be attributed to different categories of applicant covariates. We find that the greatest driver of variation in exploration bonuses is an applicant's work history variables; Black and Hispanic applicants also receive higher bonuses because they tend to have more distinctive work experiences.

A crucial question raised by this analysis is whether these differences in diversity are associated with differences in applicant quality. We will discuss this extensively in the next section and provide evidence that, despite their demographic differences, hiring outcomes for applicants selected by our SL and UCB models are comparable to each other, and much better than those selected by human recruiters.

3.5.2 UCB and SL versus Human Recruiters: Quality of Selected Applicants

While we observe demographics for all applicants, we only observe hiring potential H for applicants who are actually interviewed. We therefore cannot directly observe hiring potential for applicants selected by either algorithm, but not by the human reviewer. To address this, we take three complementary approaches, described previously in Section 3.3. Across all three approaches, we find evidence that both SL and UCB models would select applicants with greater hiring potential, relative to human screening.

Interviewed sample

Our first approach restricts to the sample of applicants who are interviewed, for whom we directly observe hiring outcomes. Among this set, we directly observe ML scores s^{SL} and s^{UCB}. We do not, however, directly observe the implicit score that human recruiters give each candidate. Before continuing, we therefore need to generate an estimate of "s^H," an applicant's propensity to be selected for an interview by a human recruiter. To do this, we simply generate a model of $E[I|X]$ where $I \in \{0, 1\}$ are realized human interview outcomes, using same logistic LASSO approach described in Section 3.4.1.[27] Appendix Figure A.8 plots the ROC associated with this model. Our model ranks a randomly chosen interviewed applicant ahead of a randomly chosen applicant who is not interviewed 76% of the time.[28]

[27]The only methodological difference between this model and our SL model is that, because we are trying to predict interview outcomes as opposed to hiring outcomes conditional on interview, our training sample consists of all applicants in the training period, rather than only those who are interviewed.

[28]Although a "good" AUC number is heavily context specific, a general rule of thumb is that models with an AUC in the range of $0.75 - 0.85$ have acceptable discriminative properties, depending on the specific context and shape of the curve (Fischer et al., 2013).

Figure 3.2 plots a binned scatterplot depicting the relationship between algorithm scores and hiring outcomes among the set of interviewed applicants; each dot represents the average hiring outcome for applicants in a given scoring ventile. Among those who are interviewed, applicants' human scores are uninformative about their hiring likelihood; if anything this relationship is slightly negative.[29]

In contrast, all ML scores have a statistically significant, positive relation between algorithmic priority selection scores and an applicant's (out of sample) likelihood of being hired.[30]

Table 3.2 examines how these differences in scores translate into differences in interview policies. To do so, we consider "interview" strategies that select the top 25, 50, or 75% of applicants as ranked by each model; we then examine how often these policies agree on whom to select, and which policy performs better when they disagree. Panel A compares the SL model to the human interview model and shows that the human model performs substantially worse in terms of predicting hiring likelihood when the models disagree: only 5-8% of candidates favored by the human model are eventually hired, compared with 17-20% of candidates favored by the SL model. Panel B finds similar results when comparing the human model to the UCB model. Finally, Panel C shows that, despite their demographic differences, the SL and UCB models agree on a greater share of candidates relative to the human model, and there do not appear to be significant differences in overall hiring likelihoods when they disagree: if anything, the UCB model performs slightly better.

For consistency, Appendix Figure A.10 revisits our analysis of diversity using the same type of selection rule described in this section: specifically, picking the top 50% of candidates among the set of interviewed. Again, we find that UCB selects a substantially more diverse set of candidates than the SL model.

Full sample, assuming no selection on unobservables

A concern with our analysis on the $I = 1$ sample is that human recruiters may add value by screening out particularly poor candidates so that they are never observed in the interview sample to begin with. In this case, then we may see little relation between human preferences and hiring potential among those who are interviewed, even though human preferences are highly predictive of quality in the full sample.

Using this approach, Figure 3.3 again shows that ML models outperform human recruiting practices. Among those selected by human recruiters, the average observed hiring likelihood is 10 percent. In contrast, our calculations show that ML models select applicants with almost 3 times higher predicted hiring potential. In particular, the average expected hiring likelihood for applicants selected by the UCB model is 27 percent and 32 percent for the SL model. The slightly weaker performance of the UCB model may be explained by

[29] This weak relation between human preferences and outcomes is consistent with existing work documenting that humans often have incorrect perceptions of worker quality. For instance, Hoffman, Kahn and Li (2017) find that firms see worse hiring outcomes when humans make exceptions to algorithmic suggestions. In a study of personnel assessment, Yu and Kuncel (n.d.) find that the scores of expert human resource managers were at best very weakly related to on the job performance.

[30] Appendix Table A.2 shows these results as regressions to test whether the relationships are statistically significant.

the fact that an emphasis on exploration means that the UCB algorithm may select weaker candidates, particularly in earlier periods. Together, this set of results are consistent with our findings from the interviewed-only subsample: the hiring yield of ML algorithms are similar to each other and in all cases better than the human decision-maker. We find no evidence that the gains in diversity that we document in Section 3.5.1 come at the cost of substantially reducing hiring rates among selected applicants.

We note that our analysis above relies on a common support assumption to form our reweighted estimates: intuitively, we are only able to infer the quality of the ML-selected applicant pool from the set of human-selected applicants if the candidates that are selected by the ML have some non-zero probability of being selected by human recruiters. Appendix Figure A.11 plots the distribution of a candidate's estimated propensity to be selected by a human recruiter, for the set of applicants chosen by SL and UCB models. In both cases, we find that all ML-selected applicants have a human selection propensity strictly between 0 and 1; we see no mass at or near zero.

Testing for selection on unobservables

Appendix Figure A.13 plots the distribution of jackknife interview pass rates in our data, restricting to the 54 recruiters (two thirds of the sample) who evaluate more than 50 applications (the mean in the sample overall is 156). After controlling for job family, job level, and work location fixed effects, the 75th percentile screener has a 50% higher pass rate than the 25th percentile screener. Appendix Table 3.3 shows that this variation is predictive of whether a given applicant is interviewed, but is not related to any of the applicant's covariates.

We are also concerned about violations of monotonicity: a lenient screener may have a different preference ordering of applicants, relative to a strict screener. We examine this in two ways. First, following the literature, e.g. Dobbie, Goldin and Yang (2018); Frandsen, Lefgren and Leslie (2019); Leslie and Pope (2017), Appendix Table A.3 shows that our leniency instrument is positively correlated with an applicant's interview likelihood across demographic, and educational groups. We also examine the preferences of lenient and strict screeners directly: using our training period sample (2016-2017), we build two models predicting an applicant's likelihood of being interviewed: one trained on data from lenient screeners and one is trained from strict screeners. In Appendix Table A.4, we show that the within-individual correlation between these two selection propensities in our analysis data (2018-2019) is high: applicants that are favored by strict reviewers are likely to be favored by lenient reviewers as well.

Figure 3.4 plots the relationship between screener leniency and interview outcomes. If humans are, on average, positively selecting candidates, then it should be the case that applicants selected by more stringent reviewers—e.g. those who are subjected to a higher bar—should be more likely to be hired conditional on being interviewed than those selected by more lenient reviewers. Panel A of Figure 3.4 shows that there does not appear to be such a relationship when we do not control for applicant covariates, indicating that, at least on the margin, humans do not necessarily interview applicants with stronger covariates. In Panel B, we introduce controls for applicant demographics and qualifications and show that there does not appear to be positive selection on unobservables either. In both panels, we

include job family, job level, and work location fixed effects to account for the possibility that interview rates may be associated with differences in hiring demand.

Marginally interviewed sample

Figure 3.5 plots the characteristics of instrument compliers with high and low ML scores, following the approach discussed in Section 3.3.2. Instrument compliers can be thought of as marginal in that they are interviewed only because they were randomly assigned to a lenient recruiter.

Panels A, C, and E focus on applicants who are marginally selected based on SL model scores while Panels B, D, and F focus on marginal applicants as defined by UCB scores. In Panels A and B, we see that compliers with high SL and UCB scores are both more likely to be hired than those with low scores. This indicates that, on the margin, nudging human interview decisions toward either ML preference would increase expected hiring yield.

In the remaining panels, we consider how following ML recommendations on the margin would change the demographics of selected candidates. In Panel C, we see that marginally selected applicants with high SL scores are substantially less likely to identify as Black or Hispanic. As such, nudging toward SL scores would tend to decrease the racial and ethnic diversity of selected applicants, relative to existing human decisions. In contrast, Panel D shows the opposite for the UCB model. Here, we find that compliers with high UCB scores are more likely to be Black or Hispanic. As such, the interview policy defined by $\tilde{I}$ would increase quality and diversity on the margin, relative to the firm's current practices. In Panels E and F, we show that both the SL and UCB models would tend to increase the representation of women.

These results are again consistent with our earlier results. In both cases, following UCB recommendations can increase hiring yield and diversity relative to the firm's present policies, while following traditional SL recommendations increases quality but decreases racial and ethnic diversity.

3.6 Alternative measures of quality

3.6.1 Maximizing Offer Rates

In our main analyses, we focus on screening models that are designed to maximize hiring yield, that is, the likelihood that a selected applicant would join the firm if interviewed. This is our preferred specification as it captures the key reason why firms turn to algorithms in the first place: the desire to fill vacancies with qualified workers.

Being hired requires that a worker both receive and accept a job offer. To isolate an algorithm's ability to identify applicants a firm would like to hire, we build an alternative set of UCB and SL models that focus on maximizing the likelihood that an applicant is extended an offer, regardless of whether they accept. These models are trained in the same way as our main models, except using offer as the outcome variable of interest. Appendix Figure A.3 shows that we correctly predict offer outcomes in our baseline training data approximately 68 percent of the time.

In Figure 3.6, we show that offer-based SL and UCB models behave similarly to our hire-based models. Panels A and B compare the demographics of applicants selected under SL and UCB models. Similar to our main results in Figure 3.1, we find that the SL model dramatically reduces the share of Black and Hispanic applicants who are selected for an interview (to less than 2% from a human recruiter baseline of just under 10%) while the UCB model increases this share to approximately 15%). In Panel C, we compare the average offer rate of UCB, SL and human selected applicants, using our inverse propensity weighting estimates discussed in Section 3.5.2.[31] Consistent with Figure 3.3, we find that both UCB and SL models outperform human recruiters, with the SL model somewhat outperforming UCB over the 18 months of our analysis period.

Figure 3.7 plots the correlation between UCB and SL scores and offer rates, among the set of applicants who are interviewed, analogous to the results presented for the hire model in Figure 3.2. Again, we find that candidates with higher UCB or SL scores are more likely to receive an offer, whereas applicants who are more likely to be interviewed by human recruiters tend to have, if anything, worse offer outcomes.

Finally, Figure 3.5 repeats our IV analysis for the offered models. In Panels A and B, we show that, among applicants who receive an interview on the margin, those with higher UCB or SL scores tend to have higher offer rates. This suggests that firms can improve offer rates by following the recommendations of either model when deciding whether or not to grant interviews on the margin. Yet, in Panels C and D, we show that following SL recommendations on the margin would significantly decrease the share of Black and Hispanic applicants who receive an interview, while following UCB recommendations would increase this share. Panels E and F show a zero or marginally negative effect of both models on the share of women who are selected. Taken together, these results again suggest that following UCB recommendations can lead firms to select a higher quality and more diverse set of applicants.

3.6.2 On the Job Performance

One concern with our analysis so far is that both hiring and offer outcomes may not be the measure of quality that firms are seeking to maximize. If firms ultimately care about on the job performance metrics, then they may prefer that its recruiters pass up candidates who are likely to be hired in order to look for candidates that have a better chance of performing well, if hired.

Our ability to assess this possibility is limited by a lack of data on tracking on the job performance. Ideally, we would like to train a model to predict on the job performance (instead of or in addition to hiring likelihood) and then compare the performance of that model to human decision-making. However, of the nearly 49,000 applicants in our training data, only 296 are hired and have data on job performance ratings, making it difficult to accurately build such a model.

We take an alternative approach and correlate measures of on the job performance with our model of human preferences, as well as our SL and UCB models, using data from our

[31] Appendix Figure A.12 looks for common support among applicants selected by a human recruiter, versus applicants chosen by SL and UCB offer models. We that all ML-selected applicants have a human selection propensity strictly between 0 and 1 with no mass at or near zero.

training period. If it were the case that humans were trading off hiring likelihood with on the job performance, then our human SL model (e.g. predicting an applicant's likelihood of being interviewed) should be positively predictive of on the job performance, relative to our ML models.

Table 3.5 presents these results using two measures of performance: on the job performance ratings from an applicant's first mid-year review, and an indicator for whether an applicant has been promoted. On the job performance ratings are given on a scale of 1 to 3, referring to below, at, or above average performance; 13% receive an above average rating. We also examine whether a worker is promoted within the time seen in our sample; this occurs for 8% of hires in the analysis period.

Panel A examines the correlation between our model of human interview behavior, our "human SL" model, and performance rating and promotion outcomes. Columns 1 examines a worker's likelihood of receiving a top performance rating and Column 2 focuses on promotions. In both cases, we observe a negatively signed and sometimes statistically significant relationship: if anything, human recruiters are less likely to interview candidates who turn out to do well on the job. In contrast, Panels B and C conduct the same exercise for each of our ML models; Columns 1 and 3 present results for models trained to maximize hiring rates and Columns 2 and 4 present results for models trained to maximize offer likelihood. For our SL hired model, these correlations are positively signed and statistically insignificant. For the SL offered model, we see a positive and statistically significant correlation between scores and top performance ratings, and a negative but statistically insignificant correlation for promotions. We find a similar pattern for the UCB scores: we see a positive and sometimes statistically significant relationships between the UCB hired model score and on the job performance. For the offered model, we again see a positive and statistically significant correlation between scores and top performance ratings, and a negative but statistically insignificant correlation for promotions

We caution that these data are potentially subject to strong sample selection—they examine the correlation between applicant scores among the 233 hires in our analysis sample, only 180 of whom have mid-year evaluation data. That said, our results provide no evidence to support the hypothesis that human recruiters are successfully trading off hiring likelihood in order to improve expected on the job performance among the set of applicants they choose to interview.

3.7 Alternative Policies

So far, we have given our algorithms access to applicant's demographics and have made no restrictions on which applicants it can select. In this section, we compare our baseline UCB analysis to two alternative approaches that treat demographic information differently. The first regulates algorithmic *inputs*: it restricts the model's ability to access information on race, gender, and ethnicity but does not restrict its choices. The second approach, in contrast, regulates algorithmic *outputs*: it maintains access to demographics, but imposes a quota on the racial shares of applicants that the model can select.

These policies can be thought of as holding up principles of discrimination law. In the US, the Equal Opportunity Employment Commission (EEOC) looks for "disparate treat-

ment" (treating applicants differently on the basis of their demographics, which race-aware algorithms do) or "disparate outcomes" (selecting a set of applicants who deviate substantially from the make-up of the applicant pool). This is similar to the European Union's Equal Treatment Directive, which prohibits "direct discrimination" (e.g. treating applicants differently on the grounds of a protected category) and "indirect discrimination" (e.g. when on their face neutral policies put certain groups at a disadvantage in hiring). Blinding algorithms are a way of preventing disparate treatment by regulating algorithmic inputs while quotas are a way of preventing disparate outcomes by regulating algorithmic outputs.

3.7.1 Demographics Blinding

Our main algorithms are trained on a variety of applicant characteristics, including explicit information on race, ethnicity, and gender. As a result, these models can treat applicants differently on the basis of protected categories, a legal area (Kleinberg et al., 2018*b*). It therefore natural to ask how our results would change if we eliminated the use of race and gender as model inputs.[32]

The theoretical impact of demographics blinding is difficult to predict because demographics information enters the baseline UCB model in two ways: first, as features of the model that are used to predict an applicant's chances of being hired if interviewed; and second, as inputs into how exploration bonuses are assigned. Eliminating this information can therefore shift the model's predictive abilities, as well as the type of exploration it engages in. For example, blinding forces the model to assign the same return to education for applicants of different ethnic groups, even though this may not be true in practice. It may also hinder the algorithm's ability to assign higher bonuses to members of racial minorities on average—but whether this is the case or not depends on whether those minorities are under-represented on other dimensions (such as education or work history) that the model can still use.

To examine what occurs in our setting, we re-estimate the UCB model without the use of applicants' race, gender, and ethnicity in either prediction or bonus provision. As a practical matter, we continue to allow the inclusion of other variables, such as geography, which may be correlated.

Figure 3.9 shows how this blinding impacts diversity. Panel A reproduces the composition of applicants selected by the unblinded UCB model and Panel B displays the blinded results. Blinding reduces the share of selected applicants who are Black or Hispanic, from 24% to 14%, although there is still greater representation relative to human hiring (10%). The most stark differences come in the treatment of White and Asian applicants. In the non-blinded model, White and Asian applicants make up a similar share of interviewed applicants (33% and 43%, respectively), even though there are substantially more Asian applicants in the overall pool. When the algorithm is blinded, however, many more Asian applicants are selected relative to White applicants (63% vs. 23%, recalling that Asian and White applicants make up 57% and 30% of the applicant pool at large, respectively).

[32]A number of recent papers have considered the impacts of anonymizing applicant information on employment outcomes (Agan and Starr, 2018; Alston, 2019; Åslund and Skans, 2012; Behaghel, Crépon and Le Barbanchon, 2015; Craigie, 2020; Doleac and Hansen, 2020; Goldin and Rouse, 2000; Kolev, Fuentes-Medel and Murray, 2019).

In our data, this likely arises for two reasons. First, Asian applicants are more likely to have an advanced degree, a trait that is more strongly rewarded for White applicants in the unblinded model: blinding the algorithm to race therefore increases the relative returns to education among Asian applicants, relative to race-aware. Second, in the race-aware model, Asian applicants received smaller exploration bonuses because they comprised a majority of the applicant pool; when bonus provision is blinded, exploration bonuses for Asian applicants increase because they are more heterogeneous on other dimensions (such as being from different countries or having niche majors) that lead to higher bonuses.

Panel C of Figure 3.9 examines the predictive accuracy of blinded vs. unblinded UCB, using the reweighting approach described in Section 3.3.2. We find that blinding leads to a small, modest decline in in the quality of algorithmically selected candidates; both models continue to substantially outperform human evaluators. In our setting, the small difference in outcomes between the blinded and unblinded UCB models likely combines two distinct impacts. First, blinding reduces the predictive ability of our models. At the same time, Asian applicants tend to have relatively higher hire rates in our data so that, in our case, blinding shifts exploration toward a higher yield group.

3.7.2 Supervised Learning with Quota

Our UCB model does not place inherent value on demographic diversity. An alternative approach to achieving greater representation is to introduce diversity as an explicit constraint. In this section, we examine the quality of applicants who would be selected if using a traditional supervised learning model, but imposing a quota that requires the demographics of selected applicants to reflect that of the applicant pool.

We implement a simple version of this idea: for every 100 applicants we see, we score applicants according to our baseline SL model but ensure that the proportion of Asian, White, and Black or Hispanic applicants selected mirrors our overall applicant pool mean as closely as possible (in our data, this is 58% Asian, 29% White, and 13% Black or Hispanic). Panels A and B of Figure 3.10 compare the demographics of candidates selected under our baseline UCB model, as well as SL with quota. As expected, the composition of applicants selected under our quota model is similar to that of the overall applicant pool (Panel A of Figure 3.1). We note that our percentages are not exact because we are working with small discrete numbers so that it is not always possible for the share of selected applicants to equal the population share.

In Panel C, we show that the quality of workers hired under the SL with quota model is substantially worse than our baseline UCB model (as well as that of our baseline SL model from Figure 3.6), and is more comparable to our baseline human model. We believe that this is due to the fact that a quota model substantially constrains the firm in terms of when it needs to select a demographically diverse candidate. To see, this Appendix Figure **??** plots the average number of Black or Hispanic applicants selected to be interviewed over our analysis period for our UCB model (dotted blue line) and our SL with quota (solid blue line). The UCB model selects more Black and Hispanic applicants on average but varies significantly in the number it selects each period. By contrast, the quota model is restricted to selecting on average one such applicant each period—no more, no less. Such constraints could reduce the quality of selected applicants of any race by selecting too many

minority candidates when their quality is low, and too few when their quality is high. Any implementation of a quota-based model would require algorithm designers to place some kind of ex-ante guidance on how many members of each group to select over a given period, resulting in some version of this problem.

3.8 Additional Results: Time Dynamics and Learning

Applicant Data: Evolution of demographics and quality

Our main results show that our UCB algorithm increases the hiring yield of selected applicants, while also increasing demographic diversity. A key question relates to how these patterns evolve over time. For example, one may be concerned that the increases in diversity we document are transient: the UCB model initially explores by selecting demographically diverse candidates, but if these candidates have poor outcomes, then the algorithm would gradually select fewer such candidates over time. It may also be the case that the UCB algorithm may perform particularly poorly in early periods as it is exploring.

In Appendix Figure A.14, we show how demographics and hiring yield evolve over time in our data. In Panel A, we find that, while noisy, there is no discernible downward trend in the proportion of Black and Hispanic candidates selected by our UCB model over time. This suggests that, in our sample, hiring outcomes for minority applicants are high enough that our models do not update downward upon selecting them. As discussed in Section 3.4.2, one may be concerned that the stability of our demographic results represents a failure to learn due to biases arising from sample selection. Our IV analysis in Section 3.5.2 provides evidence against this possibility driving our results.

Simulated Applicant Data: Changes in applicant quality

Our analysis of time dynamics is limited by sample size and timing: our analysis period spans just under 1.5 years, and we only observe hiring outcomes among candidates interviewed during this period. Combined, this give us limited opportunities to observe how our models may evolve over longer periods, or respond to substantial changes in applicant quality.

To further explore how our UCB models behave, we conduct simulations in which we change the quality of applicants who enter our analysis sample in 2018. We assign simulated values of Y_{it} in the following manner: in each simulation, we choose one racial group, R, to experience an increase (or decrease) in hiring likelihood. At the start of 2018, we assume that group R applicants have the same average hiring likelihood as their true 2018 mean.[33] Over the course of 2018, we assume that their quality linearly increases from there so that, by the end of 2018, all incoming group R candidates have $Y_{it} = 1$. In the meantime, we hold the quality of applicants from all other groups constant at their true 2018 mean. We assign values of hiring potential to *all* applicants, regardless of whether they are interviewed in practice. In this way, our simulation comes closer to a live-implementation in which we

[33]Because Y_{it} is binary, we accomplish this by sampling from a binomial distribution with the given mean we are seeking to reproduce.

would be able to update our model with the hiring outcomes of all applicants selected by our models.

To assess our models' ability to learn about changes in applicant quality, we consider how they would evaluate the *same* cohort of candidates at different points in time. Specifically, we take the actual set of candidates who applied between January 2019 and April 2019 (hereafter, the "evaluation cohort"), and estimate their ML model scores at different points in 2018. By keeping the evaluation cohort the same, we are able to isolate changes in the algorithm's scores that arise from differences in learning and exploration over time, rather than from differences in the applicant pool.

For intuition, consider the scores of candidates on January 1, 2018, the first day of the analysis period. In this case, both the SL and UCB algorithms would have the same beliefs about the hiring potential of candidates in the evaluation cohort, because they share the same estimate of $E[Y_{it}|X'_{it}; D_0]$ trained on the initial data D_0. The UCB model, however, may have a different score, because it also factors in its exploration bonus. On December 31, 2018, however, the SL and UCB algorithms would have both different beliefs (based on their potentially different history of selected applicants) and different scores (because the UCB factors in its exploration bonus in addition to expectations of quality). To better understand how the UCB model differs from the the SL, we also consider a third variant, which tracks who the UCB model would have selected based on its estimates of $E[Y_{it}|X'_{it}; D_t^{UCB}]$ alone; this model allows us to track the evolution of the UCB model's beliefs separately from its exploration behavior.

Panel A of Figure 3.11, plots the share of Black applicants who are selected in the simulation where we increase the hiring potential of Black applicants in the manner described above. We report the results of three different selection criteria. The blue dashed line reports the selection decisions of the UCB model. The UCB model rapidly increases the share of Black candidates it selects. To better understand why this happens, we plot a green dash-dot-dot line, which tracks the UCB model's *beliefs*: that is, the share of Black applicants it would select if its decisions were driven by the $\hat{E}[Y_{it}|X'_{it}; D_t^{UCB}]$ component of Equation (3.6) only, leaving out the exploration bonus. Initially, the blue dashed line is above the green dash-dot-dot line; this means that the UCB model begins by selecting more Black applicants not because it necessarily believes that they have strong hiring potential, but because it is looking to explore. Over time, the green dash-dot-dot line increases as the models see more successful Black candidates and positively updates its beliefs. Eventually, the two lines cross: at this point, the UCB model has strong positive beliefs about the hiring potential of Black applicants, but it holds back from selecting more Black candidates because it would still like to explore the quality of other candidates. By the end of the simulation period, however, exploration bonuses have declined enough so that the UCB model's decisions are driven by its beliefs, and it selects almost exclusively Black candidates.

The solid blue line shows this same process using the SL model. While it is eventually able to learn about the simulated increase in the hiring prospects of Black applicants, it does so at a significantly slower rate relative to UCB. Because supervised learning algorithms focus on maximizing current predicted hiring rates, the SL model does not go out of its way to select Black candidates. As such, it has a harder time learning that these candidates are now higher quality. This is unsurprising considering Figure 3.1, which shows that SL models are very unlikely to select Black applicants. Panel B of Figure 3.11 plots the percentage of

Hispanic applicants who are selected, under the analogous simulation in which we increase their hiring potential. Our results are broadly similar although, in this case, the difference in learning speed is less stark than for Black applicants because the baseline SL model selects a higher share of Hispanic applicants.

Panels C and D of Figure 3.11 plot outcomes for Asian and White applicants, under simulations in which quality of these groups is assumed to increase, respectively. Because these groups are already well represented in our training data, our results are slightly different. Here, the SL model learns much more quickly about changes in the quality of Asian and White applicants because it selects a large number of candidates from these groups at baseline, making it easier to pick up on changes in their quality. Another feature to note in Panels C and D is that there is a large gap between the UCB model's beliefs and its selection choices: the UCB algorithm learns very quickly about increases in the quality of Asian and White applicants but does not initially select as many of these candidates. This occurs because the UCB model is hesitant to exclusively select members of a large group (White or Asian), having seen very few Black and Hispanic applicants.[34]

Figure 3.12 plots the analogous change in the quality of selected applicants over time, for each of the four simulations above. While the SL model eventually catches up in terms of quality, we see that the UCB model outperforms earlier because it is able to more quickly identify the group with improved quality. The gap in performance between the UCB and SL models is highest when the group whose quality is improving is less likely to be selected at baseline. This is because the UCB model proactively looks for applicants with rare covariates.

In Appendix Figures A.17 and A.18, we present analogous results from simulations in which we decrease quality. When we do this for Black and Hispanic applicants, the UCB's beliefs fall very quickly. However, because Black and Hispanic candidates continue to be so rare in the data, the UCB model continues to select a small number of these candidates, in order to continue exploring, even as their overall share among those selected trends down over time. This is an example of how the UCB model trades off immediate gains in hiring yield for the option value of increased learning in the future. When we do the same for White or Asian candidates, both the SL and UCB models reduce the share of such applicants that they select at approximately the same rate; this is because both models have already seen a large number of applicants from these groups.

3.9 Conclusion

This paper makes progress on understanding how algorithmic design shapes access to job opportunity. While a growing body of work has pointed out potential gains from following algorithmic recommendations, our paper goes further to highlight the role of algorithm design on the impact and potential consequences of these decision tools. In particular, we show that, by following an algorithm that prioritizes exploration, firms can identify more candidates that meet their hiring criteria, while also increasing the representation of Black and Hispanic

[34]In contrast, the UCB is much more willing to exclusively select Black or Hispanic applicants in the simulation results from Panels A and B because it already has more certainty about the quality of White or Asian applicants.

applicants. This occurs even though our algorithm is not explicitly charged with increasing diversity, and even when it is blinded to demographic inputs.

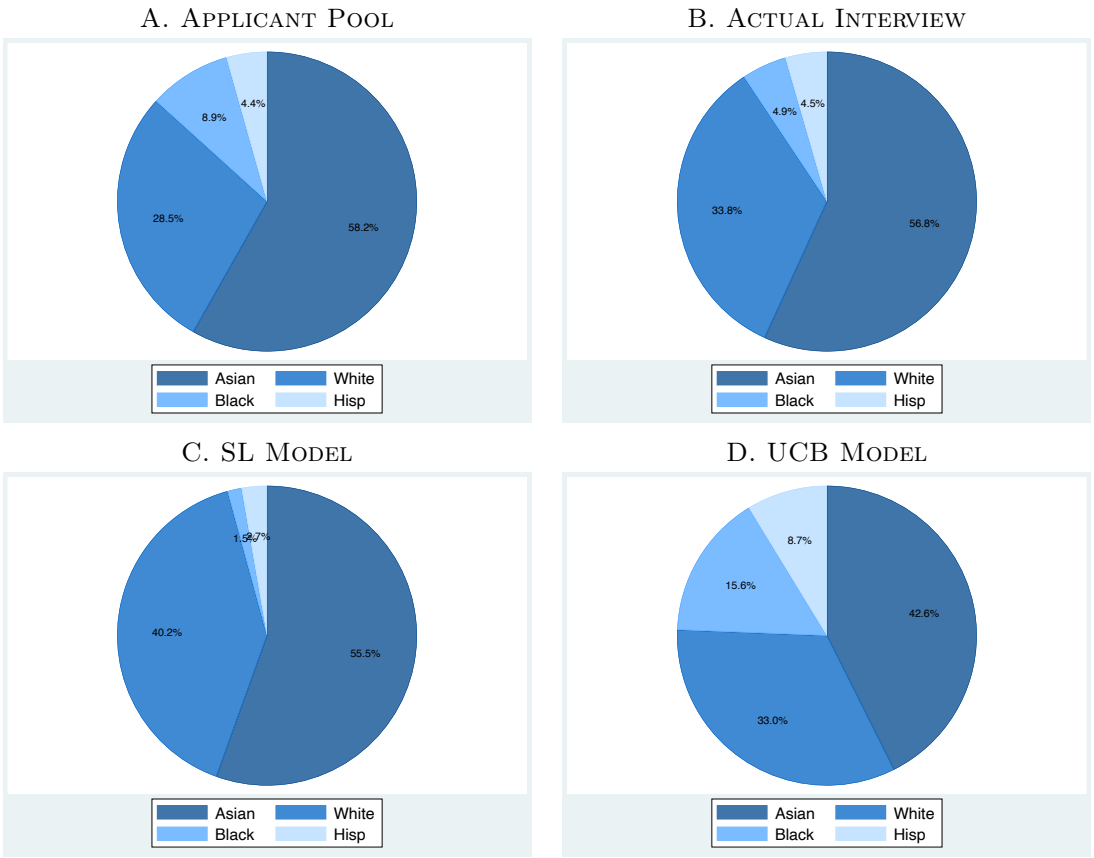
Our results shed light on the nature of the relationship between efficiency and equity in the provision of job opportunities. In our data, supervised learning algorithms substantially increase applicants' predicted hiring potential, but decrease their demographic diversity relative to the firm's actual practices. A natural interpretation of this result is that algorithms and human recruiters make different tradeoffs at the Pareto frontier, with humans placing greater value on equity at the expense of efficiency. Our UCB results, however, show that such explanations may be misleading. By demonstrating that an algorithmic approach can improve hiring outcomes while also expanding representation, we provide evidence that human recruiters are operating inside the Pareto frontier: in seeking diversity (relative to our SL model), they end up selecting weaker candidates over stronger candidates from the same demographic groups. Such behavior leaves substantial room to design and adopt data-driven approaches that are better able to identify strong candidates from under-represented backgrounds.

Finally, our findings raise important directions for future research. We focus on the use of ML to hire for high skill professional services firms; the patterns we find may not fully generalize across sectors or across firms that vary in their ability or propensity to adopt ML tools.[35] More research is needed to understand how changes in the composition of a firm's workforce—say as a consequence of adopting ML tools—would impact its future productivity and organizational dynamics. For example, there is considerable debate about the impact of diversity on team performance and how changes in the types of employees may impact other firm practices.[36] Last, as firms increasingly adopt algorithmic screening tools, it becomes crucial to understand the general equilibrium labor market effects of such changes in HR practice. For example, when adopted by a single firm, an exploration-focused algorithm may identify strong candidates who are overlooked by other firms using more traditional screening techniques; yet if all firms adopt similar exploration based algorithms, the ability to hire such workers may be blunted by supply-side constraints or competition from other firms. Such shifts in the aggregate demand for skill may also have long run impacts on the supply of skills in the applicant pool and on the returns to those skills. Both the magnitude and direction of these potentially conflicting effects deserve future scrutiny.

[35]For example, our firm has a fairly rigorous data collection process: firms that do not may make different adoption decisions and have different potential returns (Athey and Stern, 1998*a*).

[36]For instance, see Reagans and Zuckerman (2001) for a discussion of the role of diversity, and, for instance, Athey, Avery and Zemsky (2000) and Fernandez, Castilla and Moore (2000) for a discussion of how changes in firm composition can shift mentoring, promotion, and future hiring patterns.

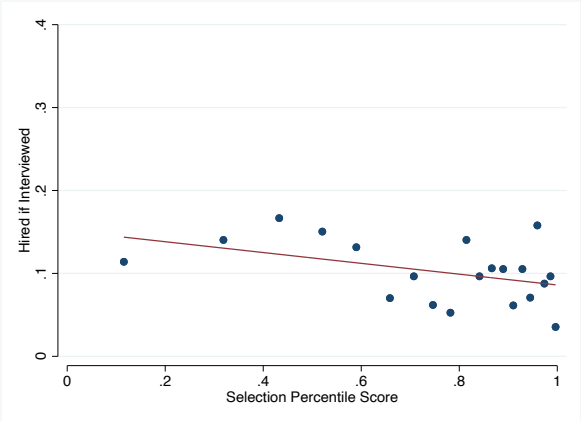
Figure 3.1: Racial Composition



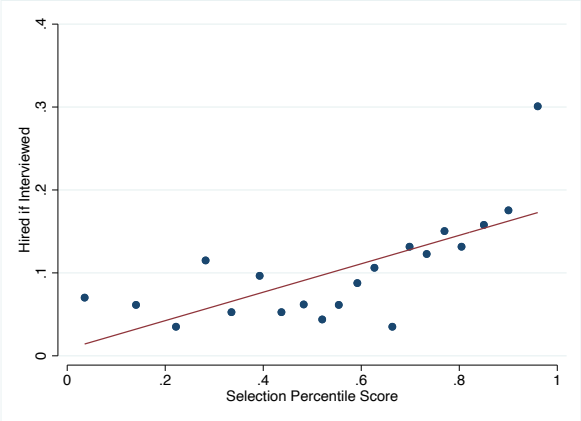
Notes: Panel A shows the racial composition of applicants in our data. Panel B shows the composition of applicants actually selected for an interview by the firm. Panel C shows the racial composition of applicants who would be selected if chosen by the supervised learning algorithm described in Equation (3.5). Finally, Panel D shows the composition of applicants who would be selected for an interview by the UCB algorithm described in Equation (3.6). All data come from the firm's application and hiring records.

FIGURE 3.2: CORRELATIONS BETWEEN ALGORITHM SCORES AND HIRING LIKELIHOOD

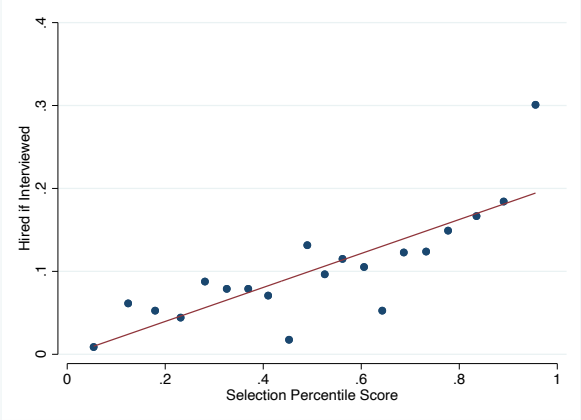
A. HUMAN



B. SL MODEL

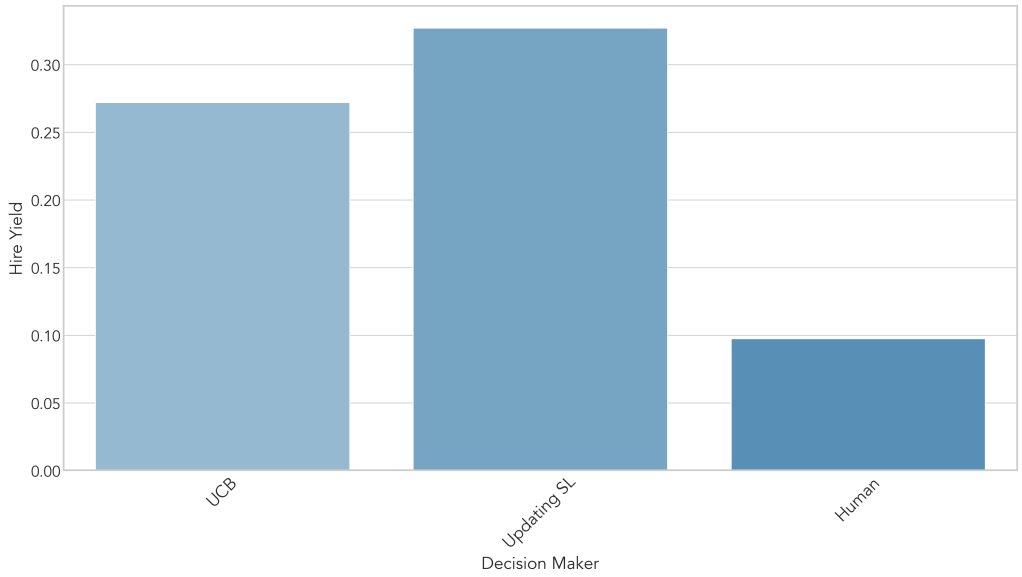


C. UCB MODEL



NOTES: Each panel of this figure plots algorithm selection scores on the x-axis and the likelihood of an applicant being hired if interviewed on the y-axis. Panel A shows the selection scores from an algorithm that predicts the firm's actual selection of which applicants to interview. Panel B shows the selection scores from the supervised learning algorithm described by Equation (3.5). Panel C shows the selection scores from the UCB algorithm described in Equation (3.6).

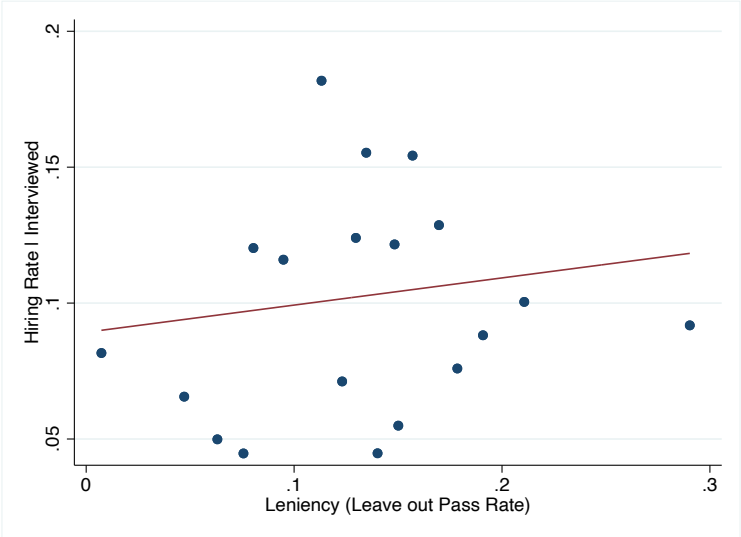
FIGURE 3.3: AVERAGE HIRING LIKELIHOOD, FULL SAMPLE



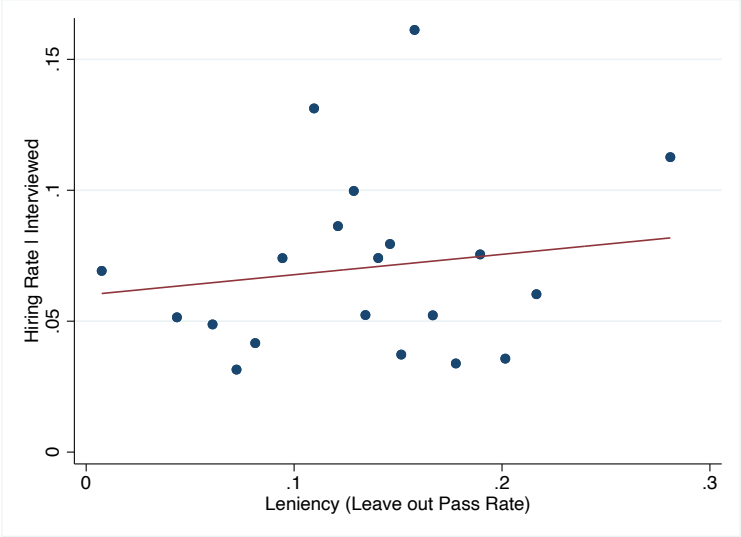
NOTES: This figure shows our inverse propensity weighting estimates of $E[Y|I^{ML} = 1]$ for each algorithmic selection strategy (SL or UCB), alongside actual hiring yields from human selection decisions.

Figure 3.4: Testing for Positive Selection

A. Full Sample, No Controls

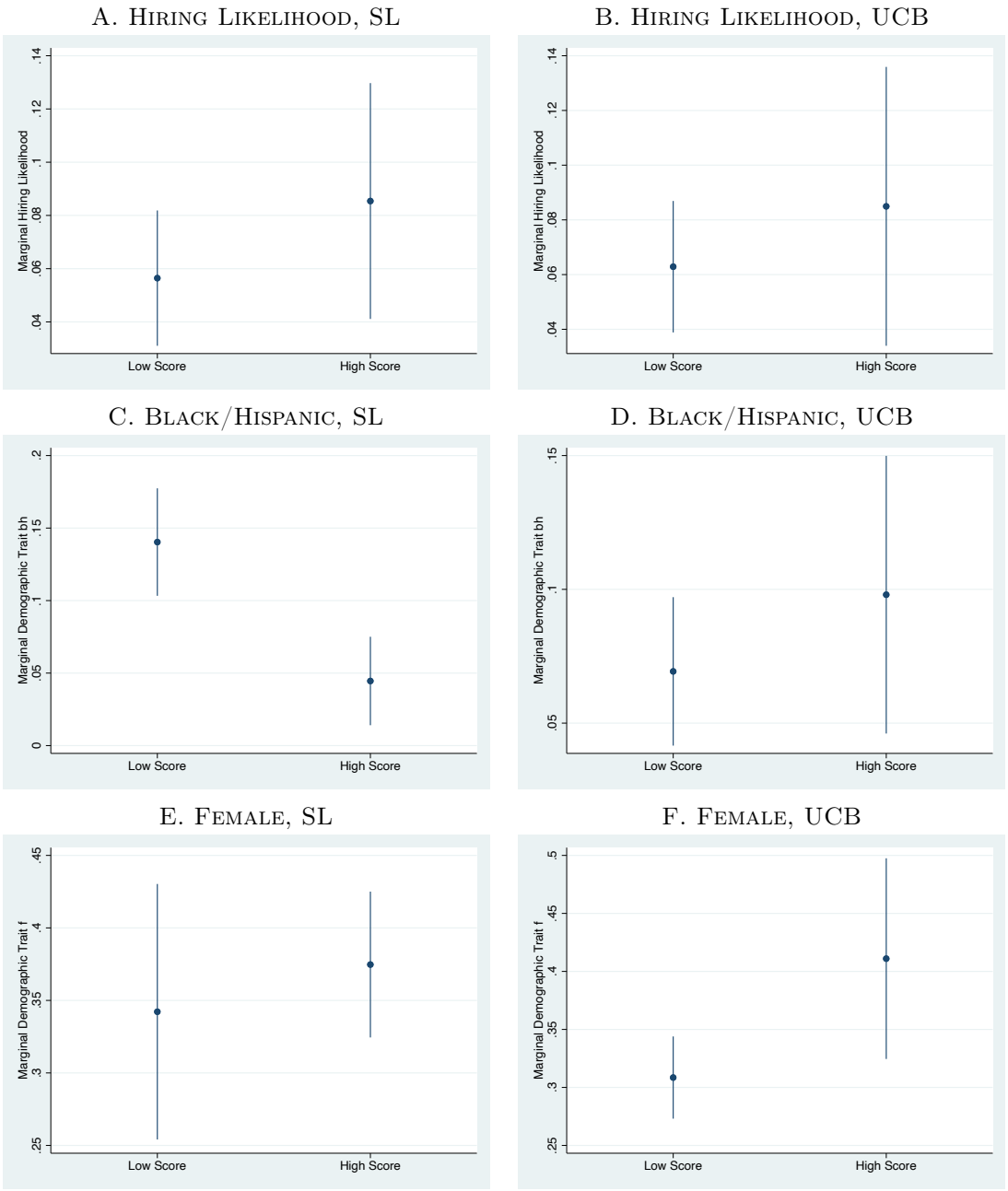


B. Full Sample, Controls



Notes: These binned scatterplots show the relationship between the leniency of randomly assigned screeners and the hiring outcomes of the applicants they select to be interviewed. Panel A plots this relationship, controlling only for job level characteristics: fixed effects for type of job, seniority level, work location, and application year. Panel B plots this relationship after adding controls for applicant characteristics: education, work history, and demographics.

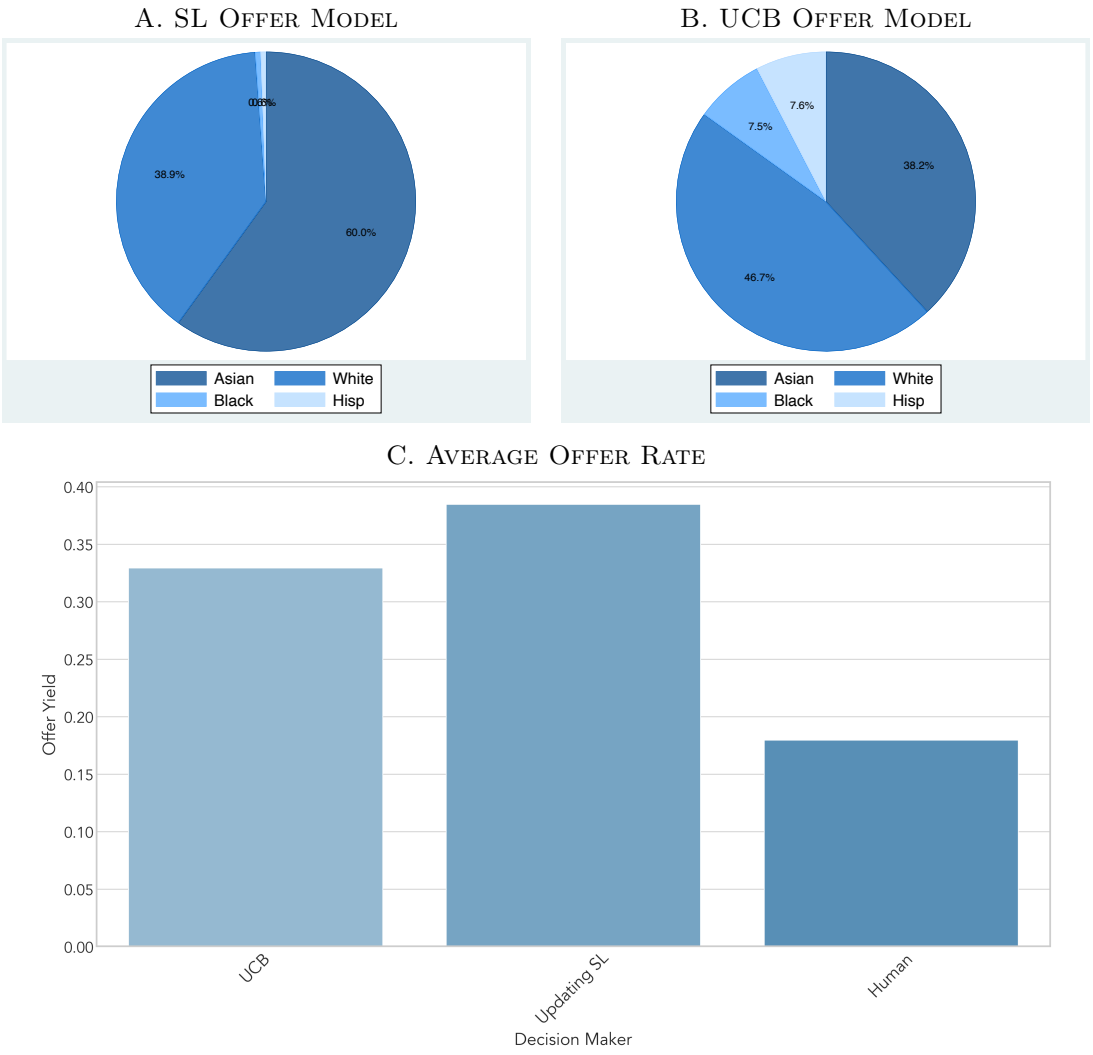
FIGURE 3.5: CHARACTERISTICS OF MARGINAL INTERVIEWEES



NOTES: NOTES: Each panel in this figure shows the results of estimating the characteristics of applicants interviewed on the margin. In each panel, these characteristics are estimated separately for applicants in the top and bottom half of the UCB algorithm's score. Panels A, C, and E consider marginal applicants as defined by SL model scores. Panels B, D, and F consider marginal applicants as defined by UCB model scores. In Panels A and B, the y-axis is the average hiring likelihood of marginally interviewed candidates; Panels C and D focus on the share of selected applicants who are Black or Hispanic; Panels E and F focus on the share of selected applicants who are female. The confidence intervals shown in each panel are derived

from robust standard errors clustered at the recruiter level.

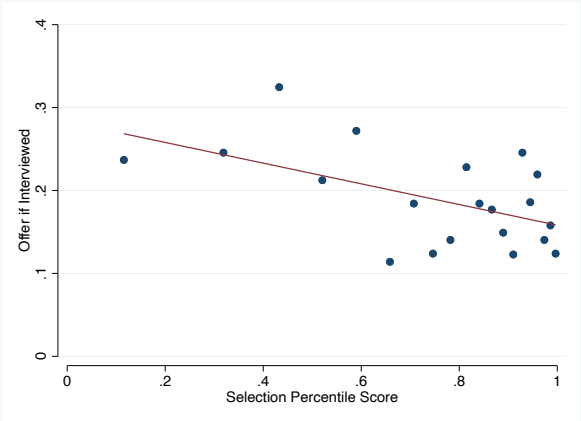
FIGURE 3.6: RACIAL COMPOSITION—OFFER MODEL



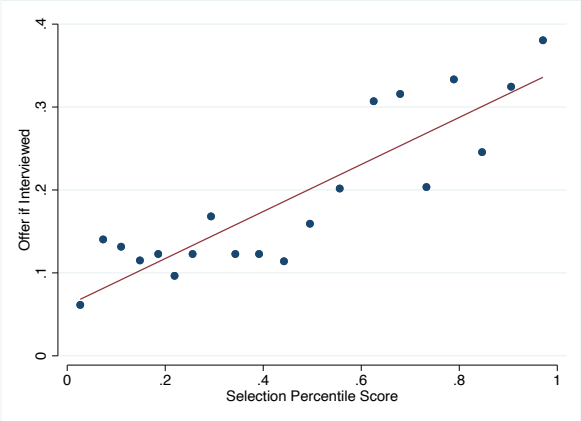
NOTES: Panel A shows the racial composition of applicants selected by an SL model trained to maximize offer likelihood. Panel B shows the composition of those who would be selected if chosen by a UCB algorithm designed to maximize offer likelihood. Panel C compares the quality (measured as the percentage of selected applicants who receive an offer using inverse propensity reweighting) for the two ML models, as well as the true offer rate from human interview decisions. All data come from the firm's application and hiring records.

FIGURE 3.7: CORRELATIONS BETWEEN ALGORITHM SCORES AND OFFER LIKELIHOOD

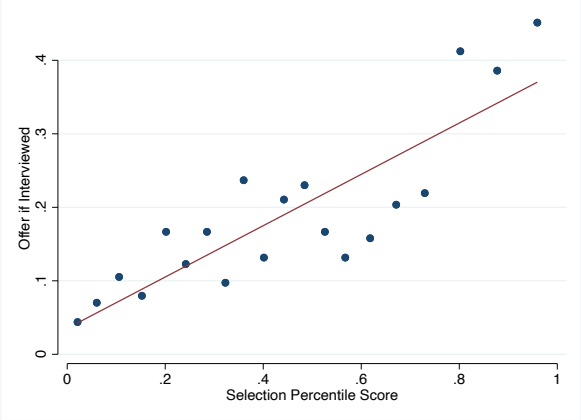
A. HUMAN



B. SL OFFER MODEL

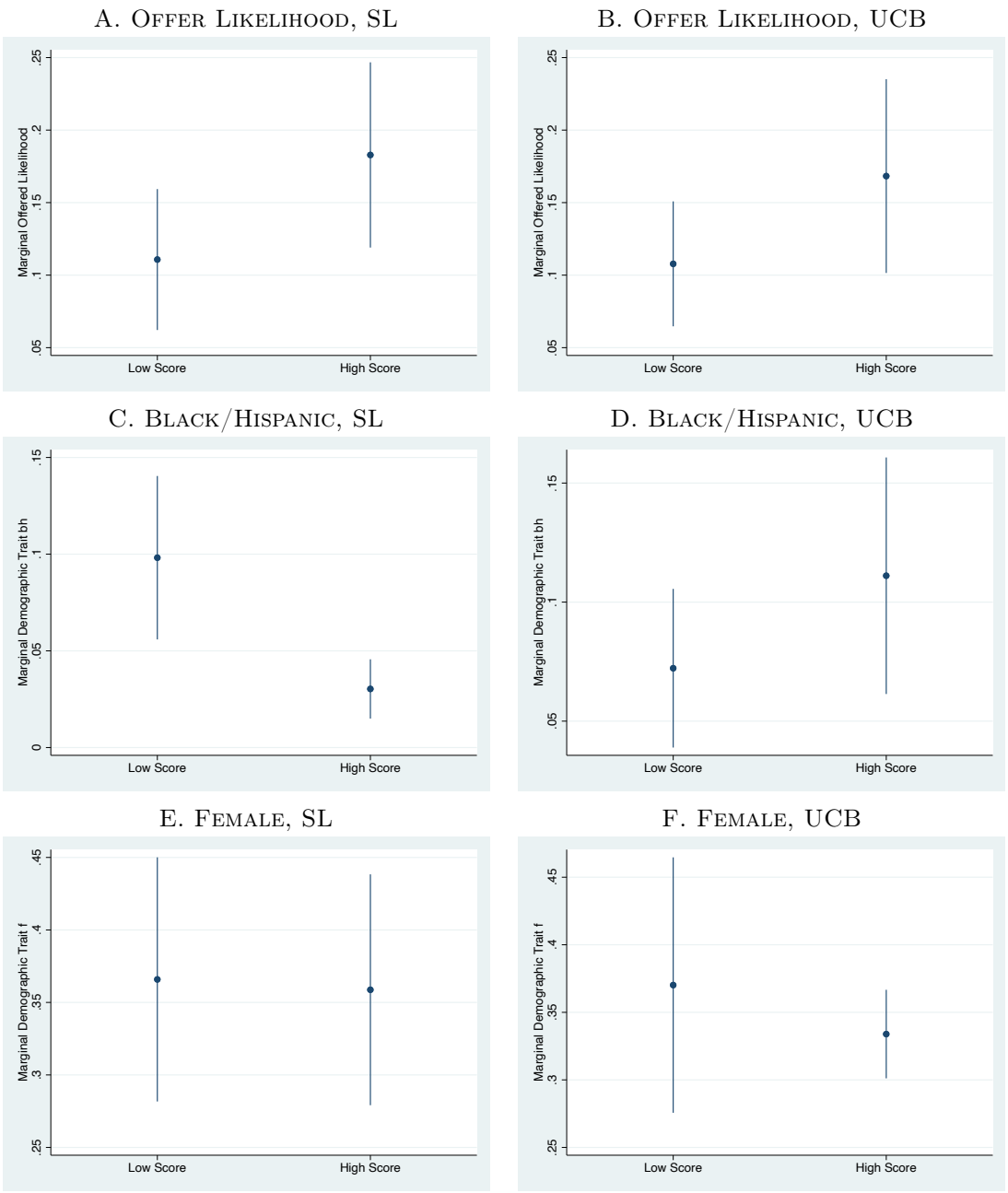


D. UCB OFFER MODEL



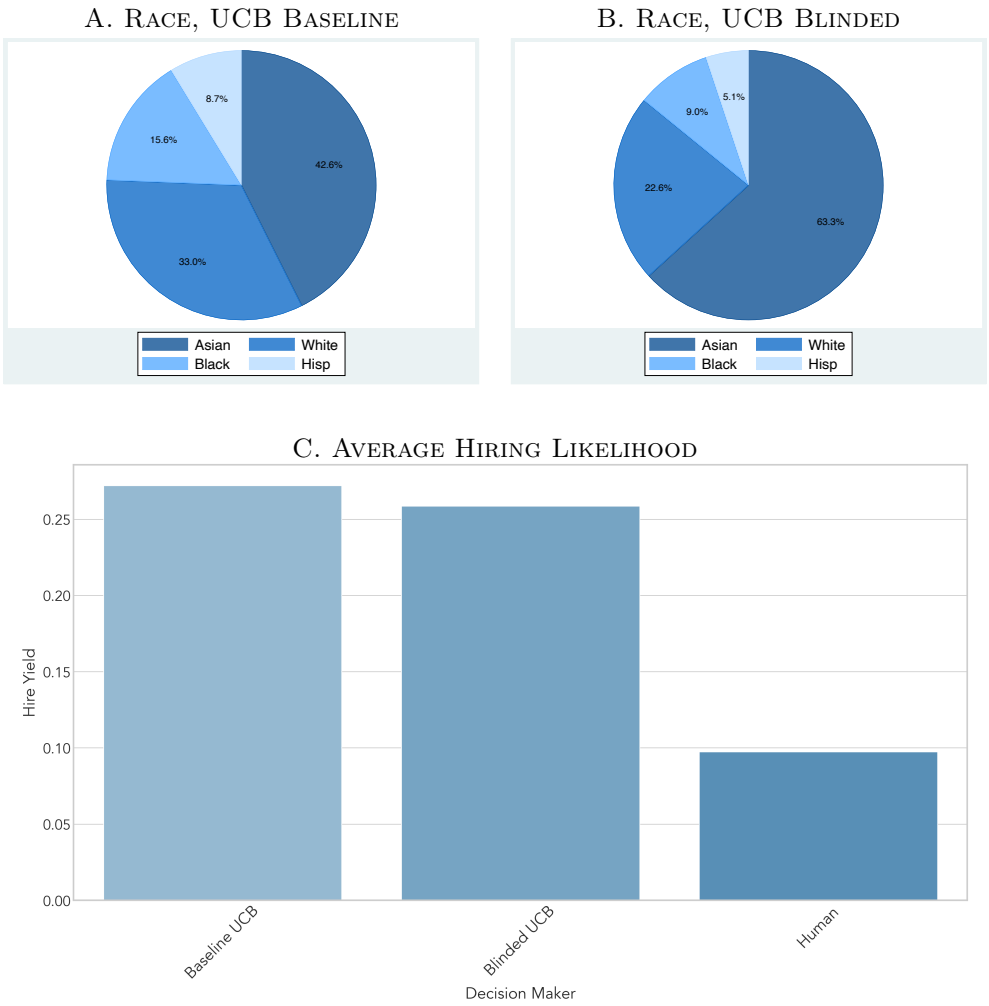
NOTES: Each panel of this figure plots algorithm selection scores on the x-axis and the likelihood of an applicant being hired if interviewed on the y-axis. Panel A shows the selection scores from an algorithm that predicts the firm's actual selection of which applicants to interview. Panel B shows selection scores from the supervised learning algorithm described in Equation (3.5). Panel C shows the selection scores from the UCB algorithm described in Equation (3.6).

FIGURE 3.8: CHARACTERISTICS OF MARGINAL INTERVIEWEES—OFFER MODELS



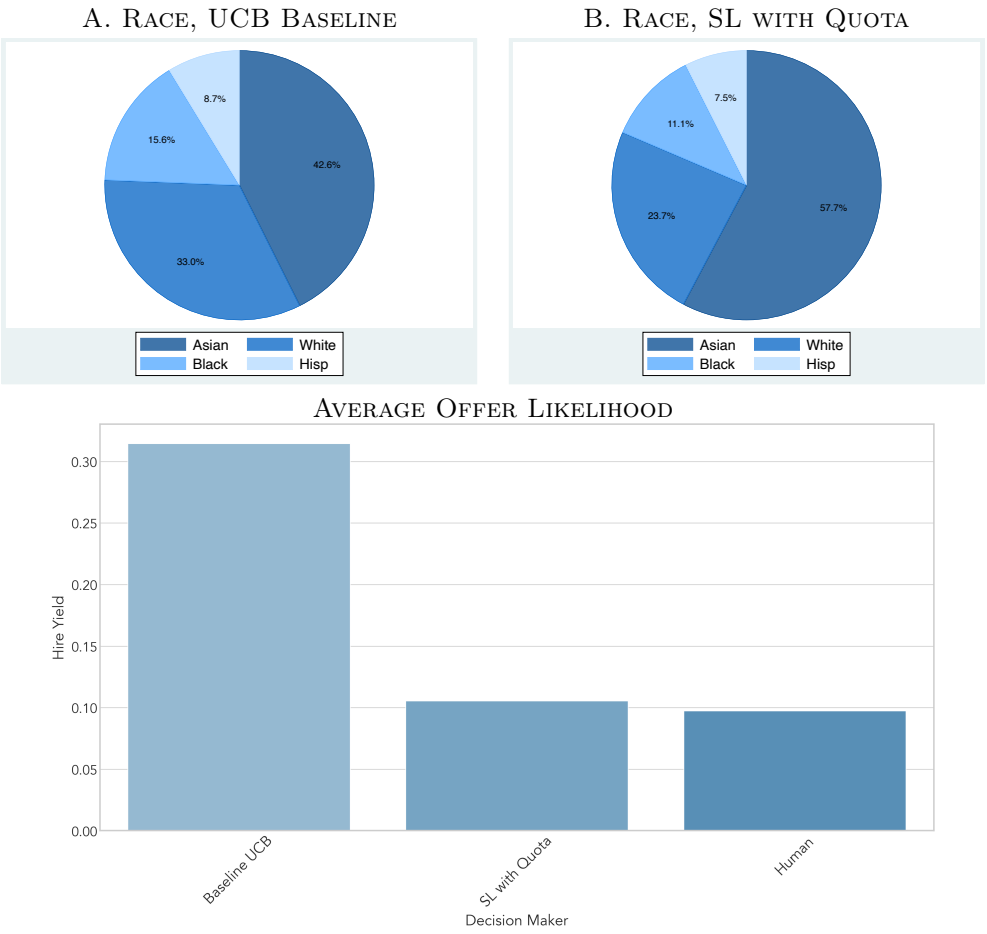
NOTES: Each panel in this figure shows the results of estimating the characteristics of applicants interviewed on the margin. In each panel, these characteristics are estimated separately for applicants in the top and bottom half of the UCB algorithm's score. Panels A, C, and E consider marginal applicants as defined by SL model scores. Panels B, D, and F consider marginal applicants as defined by UCB model scores. In Panels A and B, the y-axis is the average hiring likelihood of marginally interviewed candidates; Panels C and D focus on the share of selected applicants who are Black or Hispanic; Panels E and F focus on the share of selected applicants who are female.

Figure 3.9: Demographics Blinding



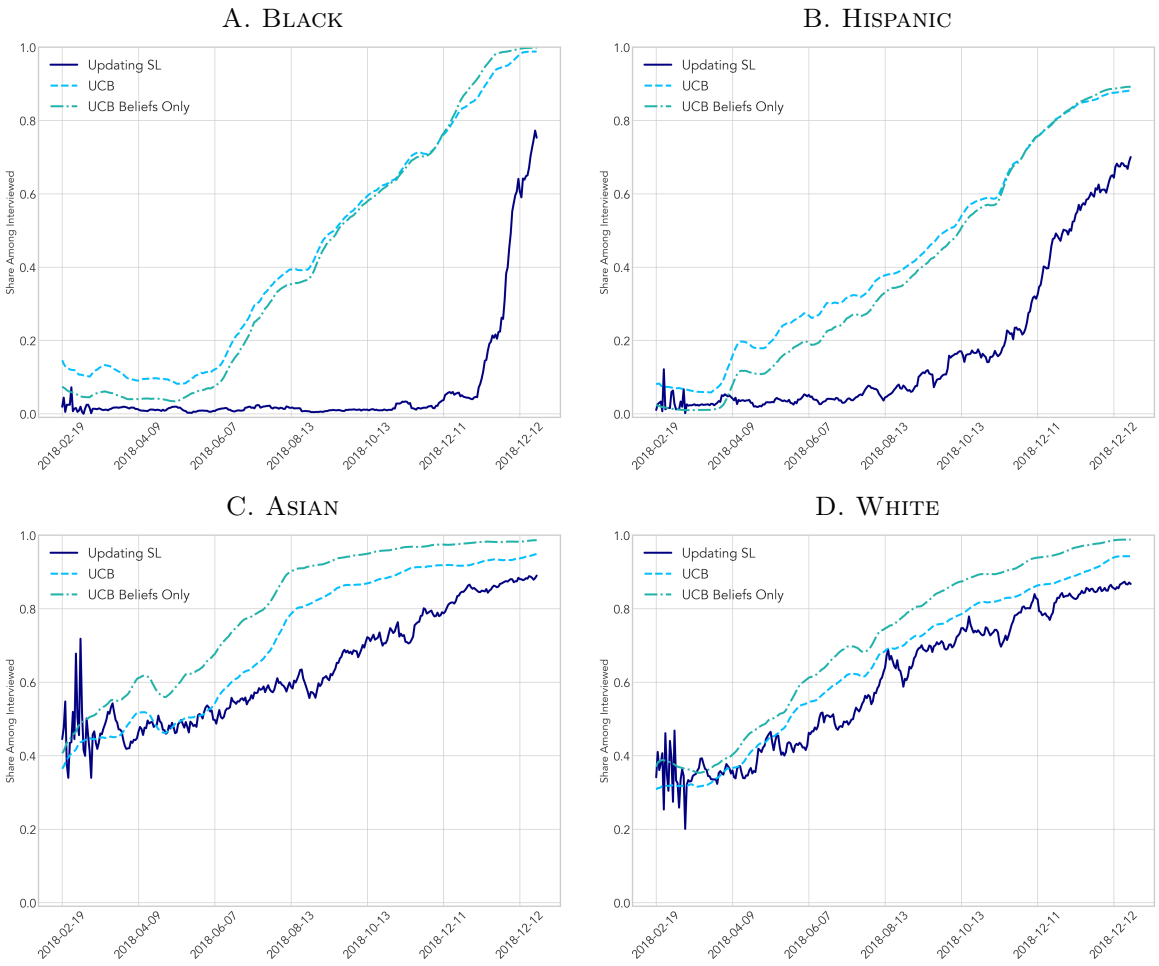
Notes: Panels A and B shows the race and gender composition of applicants recommended for interviews by the UCB algorithm when this algorithm explicitly incorporates race and gender in estimation ("baseline") and when it excludes these characteristics in estimation ("blinded"). Panel C shows our inverse propensity weighting estimates of $E[Y|I^{ML} = 1]$ for blinded vs. unblinded UCB alongside actual hiring yields from human selection decisions. All data come from the firm's application and hiring records.

FIGURE 3.10: SUPERVISED LEARNING WITH QUOTA



NOTES: Panel A shows the race and gender composition of applicants recommended for interviews by the baseline UCB algorithm. Panel B considers the alternative of using an SL model that is constrained to select applicants in proportion to their representation in the applicant pool. Panel C shows our inverse propensity weighting estimates of $E[Y|I^{ML} = 1]$ for the baseline UCB versus the SL with quota, alongside actual hiring yields from human selection decisions. All data come from the firm's application and hiring records. Panel D plots the number of Black or Hispanic applicants selected in each round for the baseline UCB and SL with quota models.

FIGURE 3.11: DYNAMIC UPDATING, INCREASED QUALITY



NOTES: This figure shows the share of applicants recommended for interviews under three different algorithmic selection strategies: SL, UCB, and the beliefs component of UCB (that is, the $\hat{E}_t[H|X; D_t^{UCB}]$ term in Equation (3.6)). In each panel, the y-axis graphs the share of "evaluation cohort" (2019) applicants who would be selected under each simulation. Panel A plots the share of evaluation cohort Black applicants who would be selected under the simulation in which the hiring potential of Black candidates increases linearly over the course of 2018, as described in Section 4.5.2. Panel B shows results from a simulation in which the hiring potential of Hispanic candidates in 2018 increases. Similarly, Panels C and D show results from simulations in which the hiring potential of White and Asian applicants increases, respectively.

FIGURE 3.12: DYNAMIC UPDATING, INCREASED QUALITY, ACCURACY

NOTES: This figure shows the share of applicants recommended for interviews that are also hired by humans under three different algorithmic selection strategies: SL, UCB, and the beliefs component of UCB (that is, the $\hat{E}_t[H|X; D_t^{UCB}]$ term in Equation (3.6)). In each panel, the y-axis graphs the share of "evaluation cohort" (2019) applicants who would be selected under each simulation who are also hired by humans. Panel A plots the evaluation cohort Black applicants who would be selected under the simulation in which the hiring potential of Black candidates increases linearly over the course of 2018, as described in Section 4.5.2. Panel B shows results from a simulation in which the hiring potential of Hispanic candidates in 2018 increases. Similarly, Panels C and D show results from simulations in which the hiring potential of White and Asian applicants increases, respectively.

TABLE 3.1: APPLICANT SUMMARY STATISTICS

Variable	Mean Training	Mean Test	Mean Overall
Black	0.087	0.087	0.087
Hispanic	0.040	0.043	0.042
Asian	0.573	0.591	0.581
White	0.300	0.279	0.290
Male	0.677	0.658	0.668
Female	0.323	0.342	0.332
Referred	0.140	0.114	0.129
B.A. Degree	0.232	0.242	0.237
Associate Degree	0.005	0.005	0.005
Master's Degree	0.612	0.643	0.626
Ph.D.	0.065	0.074	0.069
Attended a U.S. College	0.747	0.804	0.772
Attended Elite U.S. College	0.128	0.143	0.134
Interviewed	0.054	0.053	0.054
Offered	0.012	0.010	0.011
Hired	0.006	0.005	0.006
Observations	48,719	39,947	88,666

NOTES: This table shows applicants' demographic characteristics, education histories, and work experience. The sample in Column 1 consists of all applicants who applied to a position during our training period (2016 and 2017). Column 2 consists of applicants who applied during the analysis period (2018 to Q1 2019). Column 3 presents summary statistics for the full pooled sample. All data come from the firm's application and hiring records.

TABLE 3.2: PREDICTIVE ACCURACY OF HUMAN VS. ML MODELS, AMONG INTERVIEWED APPLICANTS

A. HUMAN VS. UPDATING SL

Selectivity (Top X%)	**Overlap %**	**Both**	**Human Only**	**SL Only**
	(1)	(2)	(3)	(4)
25	13.33	18.52	6.83	17.78
50	37.22	10.99	7.47	16.67
75	64.93	10.31	4.67	18.68

B. HUMAN VS. UCB

Selectivity (Top X%)	**Overlap %**	**Both**	**Human Only**	**UCB Only**
	(1)	(2)	(3)	(4)
25	15.72	17.95	6.46	20.33
50	36.00	12.09	6.33	16.57
75	61.28	10.76	3.89	16.30

C. UPDATING SL VS. UCB

Selectivity (Top X%)	**Overlap %**	**Both**	**SL Only**	**UCB Only**
	(1)	(2)	(3)	(4)
25	42.43	23.39	9.91	14.22
50	60.59	15.33	8.21	10.71
75	74.43	13.14	5.98	5.98

NOTES: This table shows the hiring rates of each algorithm when they make the same recommendation or differing recommendations. The top panel compares the human versus SL algorithm, the middle panel compares the human versus the UCB algorithm, and the lower panel compares the SL versus the UCB algorithm. Each row of a given panel conditions on selecting either the top 25%, 50%, 75% of applicants according to each of the models. For the two algorithms being compared in a given panel, Column 1 shows the percent of selected applicants that both algorithms agree on. Column 2 shows the share of applicants hired when both algorithms recommend an applicant, and Columns 3 and 4 show the share hired when applicants are selected by only one of two algorithms being compared. All data come from the firm's application and hiring records.

Table 3.3: Instrument Validity

	Interviewed	Black	Hispanic	Asian	White	Female	Ref.	MA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
JK interview rate	0.0898***	0.000158	-0.000433	0.00899	-0.00716	-0.00448	-0.0113	0.00910
	(0.00832)	(0.00470)	(0.00189)	(0.0122)	(0.00972)	(0.00557)	(0.0126)	(0.00961)
Observations	37662	37662	37662	37662	37662	37662	37662	37662

Notes: This table shows the results of regressing applicant characteristics on interviewer leniency, defined as the jack-knife mean-interview rate for the recruiter assigned to an applicant, controlling for fixed effects for job family, management level, application year and location of the job opening. This leave-out mean is standardized to be mean zero and standard deviation one. The outcome in Column 1 is an indicator variable for being interviewed. The outcomes in Columns (2)–(8) are indicators for baseline characteristics of the applicant. The sample is restricted to recruiters who screened at least 50 applicants. All data come from the firm's application and hiring records. Standard errors are clustered at the recruiter level.

Table 3.4: Impacts of following ML recommendations, IV analysis

A. Hire Rates

	Low UCB (1)	High UCB (2)	Low SL (3)	High SL (4)
Marginally Selected	0.0629***	0.0849***	0.0564***	0.0854***
	(0.0123)	(0.0260)	(0.0130)	(0.0226)
Observations	18710	18956	18862	18804

B. Offer Rates

	Low UCB (1)	High UCB (2)	Low SL (3)	High SL (4)
Marginally Selected	0.108***	0.168***	0.111***	0.183***
	(0.0220)	(0.0341)	(0.0248)	(0.0326)
Observations	18538	19128	18417	19249

C. Share Black or Hispanic

	Low UCB (1)	High UCB (2)	Low SL (3)	High SL (4)
Marginally Selected	0.0689***	0.0982***	0.139***	0.0447***
	(0.0145)	(0.0264)	(0.0190)	(0.0157)
Observations	18710	18956	18862	18804

D. Share Female

	Low UCB (1)	High UCB (2)	Low SL (3)	High SL (4)
Marginally Selected	0.309***	0.413***	0.343***	0.377***
	(0.0181)	(0.0431)	(0.0442)	(0.0252)
Observations	18710	18956	18862	18804

Notes: This table examines the characteristics of marginally interviewed applicants according to our IV strategy described in the text. Specifically, each number represents the result of the regressions outlined in Equation (3.3). The reported coefficients are the IV estimates of the coefficient on whether an applicant is interviewed and can be interpreted as the average outcome variable among treatment compliers. For example, the coefficients in Panel A Columns 1 and 2 represent the estimated average hiring rates of IV compliers with low and high UCB scores, respectively.

TABLE 3.5: CORRELATIONS BETWEEN HUMAN SCORES AND ON THE JOB PERFORMANCE

A. HUMAN SCORES

	Top Rating	Promoted
	(1)	(2)
Human SL Score	-0.282**	-0.0961
	(0.116)	(0.0782)
Observations	180	233

B. SL SCORES

	Top Rating		Promoted	
	(1)	(2)	(3)	(4)
SL Hired	0.0791		0.0816	
	(0.103)		(0.0641)	
SL Offered		0.168**		-0.0170
		(0.0800)		(0.0537)
Observations	180	180	233	233

C. UCB SCORES

	Top Rating		Promoted	
	(1)	(2)	(3)	(4)
UCB Hired	0.0377		0.161***	
	(0.106)		(0.0619)	
UCB Offered		0.163*		-0.0245
		(0.0850)		(0.0576)
Observations	180	180	233	233

NOTES: This table presents the results of regressing measures of on-the-job performance on algorithm scores, for the sample of applicants who are hired and for which we have available information on the relevant performance metric. "High performance rating" refers to receiving a 3 on a scale of 1-3 in a mid-year evaluation. Robust standard errors shown in parentheses.

Chapter 4

Generative AI at Work

Joint with Erik Brynjolfsson and Danielle Li

The emergence of generative artificial intelligence (AI) has attracted significant attention, but few studies have examined its economic impact. While various generative AI tools have performed well in laboratory settings, excitement about their potential has been tempered by concerns that these tools may be less effective in real-world settings, where they may encounter unfamiliar problems, face organizational resistance, or provide misleading information in a consequential environment (Peng et al., 2023*a*; Roose, 2023).

In this paper, we study the adoption of a generative AI tool that provides conversational guidance for customer support agents.[1] This is, to our knowledge, the first study of the impact of generative AI when deployed at scale in the workplace. We find that access to AI assistance increases the productivity of agents by 14%, as measured by the number of customer issues they are able to resolve per hour. In contrast to studies of prior waves of computerization, we find that these gains accrue disproportionately to less-experienced and lower-skill workers.[2] We argue that this occurs because generative AI systems work by capturing and disseminating the patterns of behavior that characterize the most productive agents, including knowledge that has eluded automation from earlier waves of computerization.

Computers and software have transformed the economy with their ability to perform certain tasks with far more precision, speed, and consistency than humans. To be effective, these systems typically require explicit and detailed instructions for how to transform inputs into outputs: when engineers write code to perform a task, they are codifying that task.[3] Yet because many workplace activities—such as writing emails, analyzing data, or creating presentations—rely on tacit knowledge, they have so far defied automation (Autor, 2014; Polanyi, 1966).[4]

Machine learning (ML) algorithms work differently from traditional computer programs: instead of requiring explicit instructions to function, these systems infer instructions from examples. Given a training set of images, for instance, ML systems can learn to recognize specific individuals even though one cannot fully explain what physical features characterize a given person's identity. This ability highlights a key, distinguishing aspect of ML systems: they can learn to perform tasks even when no instructions exist—including tasks requiring tacit knowledge that could previously only be gained through lived experience (Autor, 2014; Brynjolfsson and Mitchell, 2017; Polanyi, 1966).[5]

[1] A note on terminology. There are many definitions of artificial intelligence and of intelligence itself—Legg, Hutter et al. (2007) list over 70 of them. In this paper, we define "artificial intelligence" (AI) as an umbrella term that refers to systems that exhibit intelligent behavior, such as learning reasoning and problem-solving. "Machine learning" (ML) is a branch of AI that uses algorithms to learn from data, identify patterns, and make predictions or decisions without being explicitly programmed (Google, n.d.). Large language models (LLMs) and tools built around LLMs such as ChatGPT are an increasingly important application of machine learning. LLMs generate new content, making them a form of "generative AI."

[2] We provide a discussion of this literature at the end of this section.

[3] By codify, we mean compiling a process into a formal, ordered routine.

[4] Tacit knowledge refers to the knowledge and skills that individuals possess but are unable to express explicitly. It is often intuitive and nonverbal, acquired through personal experiences, observations, and practice over time. Tacit knowledge is deeply ingrained in an individual's behavior and can be difficult to transfer or convey to others through traditional methods such as training or manuals (Polanyi, 1966).

[5] As Meijer (2018) puts it "where the Software 1.0 Engineer formally specifies their problem, carefully designs algorithms, composes systems out of subsystems or decomposes complex systems into smaller components, the Software 2.0 Engineer amasses training data and simply feeds it into an ML algorithm..."

In addition, ML systems are often trained on data from human workers, who naturally differ in their abilities. By seeing many examples of tasks—making sales pitches, driving a truck, or diagnosing a patient, to name a few—performed well and poorly, these models can implicitly learn what specific behaviors and characteristics set high-performing workers apart from their less effective counterparts. That is, not only are generative AI models capable of performing complex tasks, they might also be capable of capturing the skills that distinguish top workers. The use of ML tools may therefore expose lower-skill workers to new skills and lead to differential changes in productivity.

We study the impact of generative AI on productivity and worker experience in the customer service sector, an industry with one of the highest rates of AI adoption (Chui et al., 2021). We examine the staggered deployment of a chat assistant using data from 5,000 agents working for a Fortune 500 software firm that provides business process software. The tool we study is built on a recent version of the Generative Pre-trained Transformer (GPT) family of large language models developed by OpenAI (OpenAI, 2023). It monitors customer chats and provides agents with real-time suggestions for how to respond. It is designed to augment agents, who remain responsible for the conversation and are free to ignore its suggestions.

We have three sets of findings.

First, AI assistance increases worker productivity, resulting in a 14% increase in the number of chats that an agent successfully resolves per hour. This increase reflects shifts in three components of productivity: a decline in the time it takes an agent to handle an individual chat, an increase in the number of chats that an agent handles per hour (agents may handle multiple chats at once), and a small increase in the share of chats that are successfully resolved. The productivity impacts of AI assistance are highly uneven. We find that less-skilled and less-experienced workers improve significantly across all productivity measures we consider, including a 34% increase in the number of issues they are able to resolve per hour. Access to the AI tool helps newer agents move more quickly down the experience curve: treated agents with two months of tenure perform just as well as untreated agents with more than six months of tenure. In contrast, we find minimal impacts on the productivity of more-experienced or more-skilled workers. Indeed, we find evidence that AI assistance may decrease the quality of conversations by the most skilled agents. These results contrast, in spirit, with studies that find evidence of skill-biased technical change for earlier waves of computer technology (Acemoglu and Restrepo, 2018; Autor, Levy and Murnane, 2003; Bartel, Ichniowski and Shaw, 2007; Bresnahan, Brynjolfsson and Hitt, 2002).

Our second set of results investigates the mechanism underlying our main findings. We show that AI recommendations appear useful to workers: agents who follow recommendations more closely see larger gains in productivity, and adherence rates increase over time for all workers, particularly those who were initially more skeptical. We also find that engagement with AI recommendations can generate durable learning. Using data on software outages—periods in which the AI software fails to provide any suggestions—we show that workers see productivity gains relative to their pre-AI baseline even when recommendations are unavailable. These outage-period gains are more pronounced for workers who had more prior exposure to AI assistance or who had followed AI recommendations more closely. Finally, we analyze the text of agents' chats and provide suggestive evidence that access to AI drives convergence in communication patterns: low-skill agents begin communicating more like high-skill agents.

Our third set of results focus on agents' experience of work. Work in contact centers[6] is often difficult. Agents are regularly exposed to hostile treatment from upset (and anonymous) customers, and because much work is outsourced, many agents work overnight shifts in order to service US business hours. AI assistance may help agents communicate more effectively but could also increase the likelihood that agents are perceived as mechanical or inauthentic. We show that access to AI assistance markedly improves how customers treat agents, as measured by the sentiment of their chat messages. We also find that customers are less likely to question the competence of agents by requesting to speak to a supervisor. These changes come alongside a substantial decrease in worker attrition, which is driven by the retention of newer workers.

Our overall findings show that access to generative AI can increase the productivity and retention of individual workers. We emphasize, however, that our paper is not designed to shed light on the aggregate employment or wage effects of generative AI tools. Firms may respond to increasing productivity among novice workers by hiring more of them, de-skilling positions, or seeking to develop more powerful AI systems that can replace lower-skill workers entirely. Unfortunately, our data do not allow us to observe changes in wages, overall labor demand, or the skill composition of workers hired for the job.

Our results also highlight the longer-term incentive challenges that AI systems bring. Top workers are generally not paid for their contributions to the training data that AI systems use to capture and disseminate their skills. Yet, without these contributions, AI systems may be less effective in learning to resolve new problems. Our work therefore raises questions about how workers should be compensated for the data they provide to AI systems.

Our paper is related to a large literature on the impact of technological adoption on worker productivity and the organization of work (e.g. Acemoglu and Restrepo, 2020; Acemoglu et al., 2007; Athey and Stern, 2002; Autor, Katz and Krueger, 1998; Bartel, Ichniowski and Shaw, 2007; Bloom et al., 2014; Bresnahan, Brynjolfsson and Hitt, 2002; Felten, Raj and Seamans, 2023; Garicano and Rossi-Hansberg, 2015; Hoffman, Kahn and Li, 2017; Michaels, Natraj and Van Reenen, 2014; Rosen, 1981). Many of these studies, particularly those focused on information technologies, find evidence that IT complements higher-skill workers (Akerman, Gaarder and Mogstad, 2015; Taniguchi and Yamada, 2022). Bartel, Ichniowski and Shaw (2007) shows that firms that adopt IT tend to use more skilled labor and increase skill requirements for their workers. Acemoglu and Restrepo (2020) study the diffusion of robots and find that the negative employment effects of robots are most pronounced for workers in blue collar occupations and those with less than a college education.

There have been substantially fewer studies involving AI-based technologies, generative or not. Acemoglu et al. (2022); Calvino and Fontanelli (2023); Zolas et al. (2020) examine economy-wide data from the US and OECD and show that the adoption of AI tools is concentrated among large, young firms with relatively high productivity. So far, evidence on the productivity impacts of these technologies is mixed: for example, Acemoglu et al. (2022) finds no detectable relationship between investments in AI-specific tools, while Babina et al. (2022) finds evidence of a positive relationship between firms' AI investments and their

[6]The term "contact center" updates the term "call center," to reflect the fact that a growing proportion of customer service contacts no longer involve phone calls.

subsequent growth and valuations.[7] These studies all caution that the productivity effects of AI technologies may be challenging to identify at the macro-level because AI-adopting firms differ substantially from non-adopters.

In this paper, we provide micro-level evidence on the adoption of a generative AI tool across thousands of workers working at a given firm and its subcontractors. Our work is more closely related to several other studies examining the impacts of generative AI in lab-like settings. Peng et al. (2023*b*) recruit software engineers for a specific coding task (writing an HTTP server in JavaScript) and show that those given access to GitHub Copilot complete this task twice as quickly. Similarly, Noy and Zhang (2023) conduct an online experiment showing that subjects given access to ChatGPT complete professional writing tasks more quickly. Choi and Schwarcz (2023) give law students access to AI assistance on a law school exam. Consistent with our findings, Noy and Zhang (2023), Choi and Schwarcz (2023) and Peng et al. (2023*a*) find that ChatGPT compresses the productivity distribution, with lower-skill workers benefiting the most. Our paper, however, is the first to examine longer-term effects in a real-world workplace where we can also track patterns of learning, customer-side effects and changes in the experience of work.

4.1 Generative AI and Large Language Models

In recent years, the rapid pace of AI development and public release tools such as ChatGPT, GitHub Copilot, and DALL-E have attracted widespread attention, optimism, and alarm (The White House, 2022). These technologies are all examples of "generative AI," a class of machine learning technologies that can generate new content—such as text, images, music, or video—by analyzing patterns in existing data. In this section, we provide background on generative AI as a technology and discuss its potential economic implications.

4.1.1 Technical Primer

This paper focuses on an important class of generative AI, large language models (LLMs). LLMs are neural network models designed to process sequential data (Bubeck et al., 2023). An LLM is trained by learning to predict the next word in a sequence, given what has come before, using a large corpus of text (such as Wikipedia, digitized books, or portions of the Internet). This knowledge of the statistical co-occurrence of words allows it to generate new text that is grammatically correct and semantically meaningful. Though "large language model" implies human language, the same techniques can be used to produce other forms of sequential data ("text") such as protein sequences, audio, computer code, or chess moves (Eloundou et al., 2023).

Recent progress in generative AI has been driven by four factors: computing scale, earlier innovations in model architecture, the ability to "pre-train" using large amounts of unlabeled data and refinements in training techniques.[8]

[7] OECD (2023) reports that when surveyed firms are asked directly, 57% of employers in finance and 63% in manufacturing reported that AI positively impacted productivity and 80% of surveyed workers who work with AI report higher job performance.

[8] For a more detailed technical review of progress, see Liu et al. (2023); Ouyang et al. (2022); Radford

First, the quality of LLMs is strongly dependent on scale: the amount of computing power used for training, the number of model parameters, and dataset size (Kaplan et al., 2020). Firms are increasingly devoting more resources to increasing this scale. The GPT-3 model included 175 billion parameters, was trained on 300 billion tokens, and generated approximately $5 million dollars in computing costs alone; the GPT-4 model, meanwhile, is estimated to include 1.8 trillion parameters, trained on 13 trillion tokens, at a rumored computing-only cost of $ 65 million (Brown et al., 2020; Li, 2020; Patel and Wong, 2023)

In terms of model architecture, modern LLMs use two earlier key innovations: positional encoding and self-attention. Positional encodings keep track of the order in which a word occurs in a given input.[9] Meanwhile, self-attention assigns importance weights to each word in the context of the entire input text. Together, this approach enables models to capture long-range semantic relationships within an input text, even when that text is broken up into smaller segments and processed in parallel (Bahdanau, Cho and Bengio, 2015; Vaswani et al., 2017).

Next, LLMs can be pre-trained on large amounts of unlabeled data from sources such as Reddit or Wikipedia. Because unlabeled data is much more prevalent than labeled data, LLMs can learn about natural language on a much larger training corpus (Brown et al., 2020). By seeing, for instance, that the word "yellow" is more likely to be observed with "banana" or "sun" or "rubber duckie," the model can learn about semantic and grammatical relationships even without explicit guidance (Radford and Narasimhan, 2018). The resulting model can be used in multiple applications because its training is not specific to a particular set of tasks.

Finally, general-purpose LLMs can be further "fine-tuned" to generate output that matches the priorities of any specific setting (Liu et al., 2023; Ouyang et al., 2022). For example, a model trained to generate social media content would benefit from receiving labeled data that contain not just the content of a post or tweet, but also information on the amount of user engagement it received. Similarly, an LLM may generate several potential responses to a given query, but some of them may be factually incorrect or contain toxic language. To discipline this model, human evaluators can rank these outputs to train a reward function that prioritizes desirable responses. These types of refinements can significantly improve model quality by making a general-purpose model better suited to its specific application (Ouyang et al., 2022).

Together, these innovations have generated meaningful improvements in model performance. The Generative Pre-trained Transformer (GPT) family of models, in particular, has attracted considerable media attention for their rapidly expanding capabilities.[10]

4.1.2 The Economic Impacts of Generative AI

Computers have historically excelled at executing pre-programmed instructions, making them particularly effective at tasks that can be reduced to explicit rules (Autor, 2014).

and Narasimhan (2018); Radford et al. (2019).

[9]For instance, a model would keep track of "the, 1" instead of only "the" (if "the" was the first word in the sentence).

[10]For instance, GPT-4 has recently been shown to outperform humans in taking the US legal bar exam (Bubeck et al., 2023; Liu et al., 2023; OpenAI, 2023).

Consequently, computerization has disproportionately reduced demand for workers performing "routine" tasks such as data entry, bookkeeping, and assembly line work, reducing wages in these jobs (Acemoglu and Autor, 2011). At the same time, computerization has also increased the demand for workers who possess complementary skills such as programming, data analysis, and research. Together, these changes have contributed to increasing wage inequality in the United States and have been linked to a variety of organizational changes (Autor, Levy and Murnane, 2003; Baker and Hubbard, 2003; Bresnahan, Brynjolfsson and Hitt, 2002; Katz and Murphy, 1992; Michaels, Natraj and Van Reenen, 2014; OECD, 2023).

In contrast, generative AI tools do not require explicit instructions to perform tasks. If asked to write an email denying an employee a raise, generative AI tools will likely respond with a professional and conciliatory note. This occurs because the model will have seen many examples of workplace communication in which requests are declined in this manner. Importantly, the model produces such an output even though no programmer has explicitly specified what tone would be appropriate for what context, nor even defined what a tone like "professional" or "conciliatory" means. Indeed, the ability to behave "appropriately" is one that cannot be fully articulated even by those who possess it. Rather, people learn to do so from experience and apply unconscious rules in the process. This type of "tacit knowledge" underlies most tasks humans perform, both in and out of the workplace (Autor, 2014; Polanyi, 1966).

The fact that generative AI models display such skills suggests that they can acquire tacit knowledge that is embedded in the training examples they encounter. This ability expands the types of tasks that computers may be capable of performing to include non-routine tasks that rely on judgment and experience. For example, Github Copilot, an AI tool that generates code suggestions for programmers, has achieved impressive performance on technical coding questions and, if asked, can provide natural language explanations of how the code it produces works (Nguyen and Nadi, 2022; Zhao, 2023). Meanwhile, "AI-assistant" services such as Claude can be used to produce convincing business case analyses, including reading and interpreting financial statements and offering strategic assessments. Because many of these tasks—coding, financial analysis, etc.—are currently performed by workers who have either been insulated or benefited from prior waves of technology adoption, the expansion of generative AI has the potential to shift the relationship between technology, labor productivity, and inequality (The White House, 2022).

Generative AI tools can not only expand the types of tasks that machines can perform, they may also reveal valuable information about how the most productive human workers differ from others. This is because the ML models underlying generative AI systems are commonly trained on data generated by human workers and, consequently, encounter many examples of people performing tasks both well and poorly. In learning to predict good outcomes on such data, ML models may implicitly identify characteristics or patterns of behavior that distinguish high and low performers, including subtleties rooted in tacit knowledge. Generative AI systems then take this knowledge and use it to produce new behaviors that embody what top performers might do. This ability could be used in different ways: firms may choose to replace lower-skill workers with AI-based tools, such tools could be used to demonstrate best practices to help lower-skill workers improve or help less experienced workers get up to speed more quickly. In either case, generative AI tools may have differential impacts by worker ability, even amongst workers performing the same tasks.

Despite their potential, generative AI tools face significant challenges in real-world applications. At a technical level, popular LLM-based tools, such as ChatGPT, have been shown to produce false or misleading information in unpredictable ways, generating concern about their ability to be reliable in high-stakes situations. Second, while LLM models often perform well on specific tasks in the lab (Noy and Zhang, 2023; OpenAI, 2023; Peng et al., 2023*b*), the types of problem that workers encounter in real-world settings are likely to be broader and less predictable. This raises concerns both about whether AI tools will be able to provide accurate assistance in every circumstance and—perhaps more importantly—about whether workers will be able to distinguish cases where AI tools are effective from those where they are not. Finally, the efficacy of new technologies is likely to depend on how they interact with existing workplace structures. Promising technologies may have more limited effects in practice due to the need for complementary organizational investments, skill development, or business process redesign. Because generative AI technologies are only beginning to be used in the workplace, little is currently known about their impacts.

4.2 Our Setting: LLMs for Customer Support

4.2.1 Customer Support and Generative AI

We study the impact of generative AI in the customer service industry, an area with one of the highest surveyed rates of AI adoption.[11] Customer support interactions are important for maintaining a company's reputation and building strong customer relationships, yet, as in many industries, there is substantial variation in worker productivity (Berg et al., 2018; Syverson, 2011).

Newer workers are also often less productive and require significant training. At the same time, turnover is high: industry estimates suggest that 60% of agents in contact centers leave each year, costing firms $10,000 to $20,000 dollars per agent (Buesing et al., 2020; Gretz and Jacobson, 2018). To address these workforce challenges, the average supervisor spends at least 20 hours per week coaching agents with lower performance (Berg et al., 2018). Faced with variable productivity, high turnover, and high training costs, firms are increasingly turning to AI tools (Chui et al., 2021).

At a technical level, customer support is well-suited for current generative AI tools. From an AI's perspective, customer-agent conversations can be thought of as a series of pattern-matching problems in which one is looking for an optimal sequence of actions. When confronted with an issue such as "I can't login," an AI/agent must identify which types of underlying problems are most likely to lead a customer to be unable to log in and think about which solutions typically resolve these problems ("Can you check that caps lock is not on?"). At the same time, they must be attuned to a customer's emotional response, making sure to use language that increases the likelihood that a customer will respond positively ("that wasn't stupid of you at all! I always forget to check that too!"). Because customer service conversations are widely recorded and digitized, pre-trained LLMs can be fine-tuned

[11] For instance, of the businesses that report using AI, 22% use AI in their customer service centers (Chui et al., 2021).

for customer service using many examples of both successfully and unsuccessfully resolved conversations.

Customer service is also a setting where there is high variability in the abilities of individual agents. For example, top-performing agents are often more effective at diagnosing the underlying technical issue given a customer's problem description. These workers often ask more questions before settling on a diagnosis of the problem; this takes longer initially, but reduces the likelihood that agents waste time trying to resolve the wrong problem. Such differences in agent behavior can often be inferred from the large amounts of training data that customer-service-specific AI models have access to. As a result, customer service is also a setting in which generative AI models can potentially encode some of the "best practices" that top-performing agents use.

In the remainder of this section, we provide details about the firm we study and the AI tool they adopt.

4.2.2 Data Firm Background

We work with a company that provides AI-based customer service support software (hereafter, the "AI firm") to study the deployment of their tool at one of their client firms, (hereafter, the "data firm").

Our data firm is a Fortune 500 enterprise software company that specializes in business process software for small and medium-sized businesses in the United States. It employs a variety of chat-based technical support agents, both directly and through third-party firms. The majority of agents in our sample work from offices located in the Philippines, with a smaller group working in the United States and in other countries. Across locations, agents are engaged in a fairly uniform job: answering technical support questions from US-based small business owners.

Chats are randomly assigned, and support sessions are relatively lengthy, averaging 40 minutes, with much of the conversation spent trying to diagnose the underlying technical problem. The job requires a combination of detailed product knowledge, problem solving skills, and the ability to deal with frustrated customers.

Our firm measures productivity using three metrics that are standard in the customer service industry: "average handle time," the average time an agent takes to finish a chat; "resolution rate," the share of conversations that the agent successfully resolves; and "net promoter score," (customer satisfaction), which is calculated by randomly surveying customers after a chat and calculating the percentage of customers who would recommend an agent minus the percentage who would not. A productive agent is able to field customer chats quickly while maintaining a high resolution rate and net promoter score.

Across locations, agents are organized into teams with a manager who provides feedback and training to agents. Once a week, managers hold one-on-one feedback sessions with each agent. For example, a manager might share the solution to a new software bug, explain the implication of a tax change, or suggest how to better manage customer frustration with technical issues. Agents work individually, and the quality of their output does not directly affect others. Agents are paid an hourly wage and bonuses based on their performance relative to other agents.

4.2.3 AI System Design

The AI system we study combines a recent version of GPT with additional ML algorithms specifically fine-tuned to focus on customer service interactions. The system is further trained on a large set of customer-agent conversations that have been labeled with a variety of outcomes and characteristics: whether the call was successfully resolved, how long it took to handle the call, and whether the agent in charge of the call is considered a "top" performer by the data firm. The AI firm then uses these data to look for conversational patterns that are most predictive of call resolution and handle time.

The AI firm further trains its model using a process similar in spirit to Ouyang et al. (2022) to prioritize agent responses that express empathy, provide appropriate technical documentation, and limit unprofessional language. This additional training mitigates some of the concerns associated with relying on LLMs to generate text.

Once deployed, the AI system generates two main types of output: 1) real-time suggestions for how agents should respond to customers and 2) links to the data firm's internal documentation for relevant technical issues. In both cases, recommendations are based on a history of the conversation.[12]

Figure 4.1 illustrates an example of AI assistance. In the chat window (Panel A), Alex, the customer, describes their problem to the agent. Here, the AI assistant generates two suggested responses (Panel B). In this example, it has learned that phrases like "I can definitely assist you with this!" and "Happy to help you get this fixed asap" are associated with positive outcomes. Panel A of Appendix Figure A.1 shows an example of a technical recommendation from the AI system, which occurs when it recommends a link to the data firm's internal technical documentation.

Importantly, the AI system we study is designed to augment, rather than replace, human agents. The output is shown only to the agent, who has full discretion over whether to incorporate (fully or partially) the AI suggestions. This reduces the likelihood that off-topic or incorrect outputs make their way into customer conversations. Furthermore, the system does not provide suggestions when it has insufficient training data for that situation. In these situations, the agent must respond on their own.

[12]For example, the correct response when a customer says "I can't track my employee's hours during business trips" depends on what version of the data firm's software the customer uses. Suppose that the customer has previously mentioned that they are using the premium version. In that case, they should have access to remote mobile device timekeeping, meaning that the support agents need to diagnose and resolve a technical issue that prevents the software from working. If, however, the customer stated that they are using the standard version, then the correct solution is for the customer to upgrade to the premium version in order to access this feature. For more on context tracking, see, for instance, Dunn, Inkpen and Andonie (2021).

4.3 Deployment, Data, and Empirical Strategy

4.3.1 AI Model Deployment

The AI assistant we study was gradually rolled out at the agent level after an initial seven-week randomized pilot featuring 50 agents.[13] The deployment was largely uniform across both the data firm's own customer service agents and its outsourced agents. Appendix Figure A.2 documents the progression of deployment among agents who are eventually treated. The bulk of the adoption occurs between November 2020 and February 2021.

4.3.2 Summary Statistics

Table 4.1 provides details on sample characteristics, divided into three groups: agents who are never given access to the AI tool during our sample period ("never treated"), pre-AI observations for those who are eventually given access ("treated, pre"), and post-AI observations ("treated, post"). In total, we observe the conversation text and outcomes associated with 3 million chats by 5,179 agents. Within this, we observe 1.2 million chats by 1,636 agents in the post-AI period. Most agents in our sample, 89%, are located outside of the United States, primarily in the Philippines. For each agent, we observe their assigned manager, tenure, geographic location, and firm information.

To examine the impacts of this deployment, we construct several key variables, all aggregated to the agent-month level, which is our primary level of analysis.

Our primary measure of productivity is resolutions per hour (RPH), the number of chats a worker is able to successfully resolve per hour. We consider this measure to be the most effective summary of a worker's productivity at the firm. An agent's RPH is determined by several factors: the average time it takes an agent to complete a conversation, the number of conversations they are able to handle per hour (accounting for multiple simultaneous conversations), and the share of conversations that are successfully resolved. We measure these individually as, respectively, average handle time (AHT), chats per hour (CPH), and resolution rate (RR). In addition, we also observe a measure of customer satisfaction through an agent's net promoter score (NPS), which is collected by the firm from post-call customer surveys.

We observe these measures for different numbers of agents. In particular, we are able to reconstruct measures of average handle time and chats per hour from our chat level data. We therefore observe AHT and CPH measures for all agents in our sample. Measures that involve an understanding of call quality—resolution rates, and customer satisfaction—are provided at the agent-month level by our data firm. Because our data firm outsources most of its customer service functions, it does not have direct control over this information, which is kept by subcontracted firms. As a result, we observe resolution rates and net promoter scores for a subset of agents in our data. This, in turn, means that we only observe our omnibus productivity measure—resolutions per hour—for this smaller subset.

Figure 4.2 plots the raw distributions of our outcomes for each of the never, pre-, and

[13] Data from the RCT is included as part of our primary analysis but is not analyzed separately because of its small sample size.

post-treatment subgroups. Several of our main results are readily visible in these raw data. In Panels A through D, we see that post-treatment agents do better along a range of outcomes, relative to both never-treated agents and pre-treatment agents. In Panel E, we see no discernible differences in surveyed customer satisfaction among treated and non-treated groups.

Focusing on our main productivity measure, Panel A of Figure 4.2 and Table 4.1 show that never-treated agents resolve an average of 1.7 chats per hour, whereas post-treatment agents resolve 2.5 chats per hour. Some of this difference may be due to differences in the initial section: treated agents have higher resolutions per hour prior to AI model deployment (2.0 chats) relative to never treated agents (1.7). This same pattern appears for chats per hour (Panel C) and resolution rates (Panel D): while ever-treated agents appear to be stronger performers at the outset than agents who are never treated, post-treatment agents perform substantially better. When looking instead at average handle times (Panel B), we see a starker pattern: pre-treatment and never-treated agents have similar distributions of average handle times, centered at 40 minutes, but post-treatment agents have a lower average handle time of 35 minutes. These figures, of course, reflect raw differences that do not account for potential confounding factors such as differences in agent experience or differences in selection into treatment. In the next section, we will more precisely attribute these raw differences to the impact of AI model deployment.

4.3.3 Empirical Strategy

We isolate the causal impact of access to AI recommendations using a standard difference-in-differences regression:

$$y_{it} = \delta_t + \alpha_i + \beta AI_{it} + \gamma X_{it} + \epsilon_{it} \tag{4.1}$$

Our outcome variables y_{it} capture various measures of productivity for agent i in year-month t, as outlined earlier. Because workers often work only for a portion of the year, we include only year-month observations for an agent who is actively employed (e.g. assigned to chats). Our main variable of interest is AI_{it}, an indicator equal to one if agent i has access to AI recommendations at time t. All regressions include year-month fixed effects δ_t to control for common, time-varying factors such as tax season or the end of the business quarter. In our preferred specification, we also include controls for time-invariant agent-level fixed effects α_i and time-varying agent tenure. Standard errors are clustered at the agent level.

A rapidly growing literature has shown that two-way fixed effects regressions deliver consistent estimates only with strong assumptions about the homogeneity of treatment effects, and may be biased when treatment effects vary over time or by adoption cohort (Borusyak, Jaravel and Spiess, 2022; Callaway and Sant'Anna, 2021; Cengiz et al., 2019; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021). For example, workers may take time to adjust to using the AI system, in which case its impact in the first month may be smaller. Alternatively, the onboarding of later cohorts of agents may be smoother, so that their treatment effects may be larger.

We study the dynamics of treatment effects using the interaction weighted (IW) estimator proposed in Sun and Abraham (2021). Sun and Abraham (2021) show that this estimator is

consistent assuming parallel trends, no anticipatory behavior, and cohort-specific treatment effects that follow the same dynamic profile.[14] In the appendix, we show that both our main differences-in-differences and event study estimates are similar using robust estimators introduced in de Chaisemartin and D'Haultfœuille (2020), Borusyak, Jaravel and Spiess (2022), Callaway and Sant'Anna (2021), and Sun and Abraham (2021), as well as using traditional two-way fixed effects OLS.

4.4 Main Results

4.4.1 Productivity Metrics

Table 4.2 examines the impact of the deployment of the AI model on our primary measure of productivity, resolutions per hour, using a standard two-way fixed effects model. In Column 1, we show that, controlling for time and location fixed effects, access to AI recommendations increases resolutions per hour by 0.47 chats, up 22.2% from an average of 2.12. In Column 2, we include fixed effects for individual agents to account for potential differences between treated and untreated agents. In Column 3, we include additional controls for the time-varying agent tenure. As we add controls, our effects fall slightly, so that, with agent and tenure fixed effects, we find that the deployment of AI increases RPH by 0.30 chats or 13.8%. Columns 4 through 6 produce these same patterns and magnitudes for the log of RPH.

Appendix Table A.1 finds similar results using alternative difference-in-difference estimators introduced in Callaway and Sant'Anna (2021), Borusyak, Jaravel and Spiess (2022), de Chaisemartin and D'Haultfœuille (2020), and Sun and Abraham (2021). Unlike traditional OLS, these estimators avoid comparing between newly treated and already treated units. In most cases, we find larger effects of AI assistance using these alternatives.

Figure 4.3 shows the accompanying IW event study estimates of Sun and Abraham (2021) for the impact of AI assistance on RPH, in levels and logs. For both outcomes, we find a substantial and immediate increase in productivity in the first month of deployment. This effect grows slightly in the second month and remains stable and persistent up to the end of our sample. Appendix Figure A.3 shows that this pattern can be seen using alternative event study estimators as well: Callaway and Sant'Anna (2021), Borusyak, Jaravel and Spiess (2022), de Chaisemartin and D'Haultfœuille (2020), and traditional two-way fixed effects.

In Table 4.3, we report additional results using our preferred specification with year-month, agent, and agent tenure fixed effects. Column 1 documents a 3.8 minute decrease in the average duration of customer chats, a 9% decline from the baseline mean (shorter handle times are generally considered better). Next, Column 2 indicates a 0.37 unit increase in the number of chats that an agent can handle per hour. Relative to a baseline mean of 2.6, this represents an increase of roughly 14%. Unlike average handle time, chats per hour account for the possibility that agents may handle multiple chats simultaneously. The fact that we

[14] This last assumption means that treatment effects are allowed to vary over event-time and that average treatment effects can vary across adoption-cohorts (because even if they follow the same event-time profile, we observe different cohorts for different periods of event-time).

find a stronger effect on this outcome suggests that AI enables agents to both speed up chats and multitask more effectively.

Column 3 of Table 4.3 indicates a small 1.3 percentage point increase in chat resolution rates, significant at the 10% level. This effect is economically modest, given a high baseline resolution rate of 82%; we interpret this as evidence that improvements in chat handling do not come at the expense of problem solving on average. Finally, Column 4 finds no economically significant change in customer satisfaction, as measured by net promoter scores: the coefficient is -0.13 percentage points and the mean is 79.6%. Columns 5 through 8 report these results for logged outcomes. Going forward, we will report our estimates in logs, for ease of interpretation.

Figure 4.4 presents the accompanying event studies for additional outcomes. We see immediate impacts on average handle time (Panel A) and chats per hour (Panel B), and relatively flat patterns for resolution rate (Panel C) and customer satisfaction (Panel D). We therefore interpret these findings as saying that, on average, AI assistance increases productivity without negatively impacting resolution rates and surveyed customer satisfaction.

4.4.2 Impacts by Agent Skill and Tenure

There is substantial debate about the distributional consequences of AI-based technologies on worker productivity. An extensive literature suggests earlier waves of information and communication technology (e.g., the Internet, computers, network-based communication) have complemented high-skill workers, increasing their productivity and labor demand and widening wage differentials. Generative AI tools, however, are based on machine learning tools that rely on looking for patterns associated with success. As discussed earlier, generative AI tools may have a different pattern of productivity consequences relative to earlier waves of technology adoption. In this section, we examine whether access to AI assistance has different impacts along two dimensions: worker skill and worker experience.

Pre-treatment Worker Skill

In Panel A of Figure 4.5, we consider how our estimated productivity effects differ by an agent's pre-AI productivity. We divide agents into quintiles using a skill index based on their average call efficiency, resolution rate, and surveyed customer satisfaction in the quarter prior to the adoption of the AI system. These skill quintiles are defined within a firm-month. To isolate the impact of worker skill, we also control for worker tenure at AI deployment.

In Panel A, we show that the productivity impact of AI assistance is most pronounced for workers in the lowest skill quintile (leftmost side), who see a .29 log point or 34% increase in resolutions per hour. In contrast, AI assistance does not lead to any productivity increase for the most skilled workers (rightmost side).

In Figure 4.6 we show that less-skilled agents consistently see the largest gains across our other outcomes. For the highest-skilled workers, we find mixed results: a zero effect on average handle time (Panel A), a positive effect for chats per hour (Panel B), and, interestingly, a small but statistically significant *decreases* in resolution rates and customer satisfaction (Panels C and D). These results are consistent with the idea that generative AI tools may function by exposing lower-skill workers to the best practices of higher-skill

workers. Lower-skill workers benefit because AI assistance provides them with new solutions, whereas the best performers may see little benefit from being exposed to their own best practices. Indeed, the fact that we find negative effects along measures of chat quality—resolution rate and customer satisfaction—suggests that AI recommendations may distract top performers, or lead them to choose the faster option (following suggestions) rather than taking the time to come up with their own responses.

Pre-treatment Worker Experience

Next, we repeat our previous analysis for agent tenure. To do so, we divide agents into five groups based on their tenure at the time the AI model is introduced. Some agents have less than a month of tenure when they receive AI access, while others have more than a year of experience. To isolate the impact of worker tenure, we control for worker skill when given access to the AI.

In Panel B of Figure 4.5, we see a clear, monotonic pattern in which the least experienced agents see the greatest gains in resolutions per hour. Agents with less than 1 month of tenure improve their resolutions per hour by .38 log points or 46% improvement (relative to agents of the same tenure who do not have access to AI assistance). In contrast, we see no effect for agents with more than a year of tenure.

In Figure 4.7, we show the same patterns for other outcomes. In Panels A and B, we see that AI assistance generates large gains in call handling efficiency, measured by average handle times and chats per hour, respectively, among the newest workers. In Panels C and D, we find positive impacts of AI assistance on chat quality, as measured by resolution rates and customer satisfaction, respectively. For the most experienced workers, we see modest positive effects for average handle time (Panel A), positive but statistically insignificant effects on chats per hour (Panel B), and small but statistically significant negative effects for measures of call quality and customer satisfaction (Panels C and D).

Moving Down the Experience Curve

To further explore how AI assistance impacts newer workers, we examine how worker productivity evolves on the job.[15] In Figure 4.8, we plot productivity variables by agent tenure for three distinct groups: agents who never receive access to the AI model ("never treated"), those who have access from the time they join the firm ("always treated"), and those who receive access in their fifth month with the firm ("treated 5 mo.").

We see that all agents begin with around 2.0 resolutions per hour. Workers who are never treated (blue line) slowly improve their productivity with experience, reaching approximately 2.5 resolutions per hour 8 to 10 months later. In contrast, workers who always have access to AI assistance (red line) increase their productivity to 2.5 resolutions per hour after only two months and continue to improve until they are resolving more than 3 chats per hour after five months of tenure.[16] Comparing just these two groups suggests that access to AI recommendations helps workers move more quickly down the experience curve.

[15] We avoid the term "learning curve" because we cannot distinguish if workers are learning or merely following recommendations.

[16] Our sample ends here because we have very few observations more than five months after treatment.

The final group in Panel A tracks workers who begin their tenure with the firm without access to AI assistance, but who receive access after five months on the job (green line). These workers improve slowly in the same way as never-treated workers for the first five months of their tenure. Starting in month five, however, these workers gain access and we see their productivity rapidly increase following the same trajectory as the always-treated agents. In Appendix Figure A.4, we plot these curves for other outcomes. We see clear evidence that the experience curve for always-treated agents is steeper for handle time, chats per hour, and resolution rates (Panels A through C). Panel D follows a similar but noisier pattern for customer satisfaction.

Together, these results indicate that access to AI helps new agents move more quickly down the experience curve. Across many of the outcomes in Figure 4.8, agents with two months of tenure and access to AI assistance perform as well as or better than agents with more than six months of tenure who do not have access.

4.5 Adherence, Learning, and Conversational Change

In this section, we conduct a variety of analyses aimed at better understanding the mechanisms behind our main results.

First, we examine how workers engage with AI recommendations. We show that workers are selective about the recommendations they adopt, following the recommendations 35% on average. We find that the returns to AI assistance are highest for workers who choose to follow recommendations. Consistent with a story in which workers find AI recommendations helpful, we show that adherence rates increase over time for all workers, especially among older workers: by the end of our sample, we see similar adherence rates across worker tenure and skill.

Second, we explore whether AI-assistance helps workers learn. Using information on software outages in which AI assistance is temporarily unavailable, we provide evidence that exposure to AI leads to durable changes in worker skills. We find that workers exposed to AI recommendations continue to perform better during outages, and this effect is greater after more exposure and for agents who more closely follow AI recommendations when the software is working.

Lastly, using text-based analysis of chat records themselves, we provide suggestive evidence that AI assistance changes the content of agents' communication. We document within-agent changes in communication following AI deployment, with larger changes for lower-skill workers. Across-person, we show that these changes increase the similarity of communication patterns between low- and high-skill agents. These results are consistent with AI recommendations leading lower-skill workers to communicate more like high-skill workers.

Taken together, our results suggest that examining and following AI recommendations helps workers—particularly lower-skilled workers—learn to adopt best practices gathered from higher-skill and more experienced agents.

4.5.1 Adherence to AI recommendations

The AI tool we study makes suggestions, but agents are ultimately responsible for what they say to the customer. In our main results, we estimate how access to the AI tool impacts outcomes regardless of how frequently agents follow its recommendations. Here, we examine how closely agents adhere to AI recommendations, and document the association between adherence and returns to adoption.

We measure "adherence" starting at the chat level, using the share of AI recommendations that each agent follows. Our AI firm codes agents as having adhered to a recommendation if they either click to copy the suggested AI text or if they self-input something very similar. We take this chat-level measure and aggregate it to the agent-month level.

Panel A of Figure 4.9 shows the distribution of average agent-month-level adherence for our post-AI sample, weighted by the log number of AI recommendations provided to that agent in that month. The average adherence rate is 38%, with an interquartile range of 23% to 50%: agents frequently ignore recommendations. In fact, the share of recommendations followed is similar to the share of other publicly reported numbers for generative AI tools; a study of GitHub Copilot reports that individual developers use 27% to 46% of code recommendations (Zhao, 2023). Such behavior may be appropriate, given that AI models may make incorrect or irrelevant suggestions.

Panel B of Figure 4.9 shows that *returns* to AI model deployment are higher when agents actually follow recommendations. To show this, we divide agents into quintiles based on the percent of AI recommendations they follow in the first month of AI access and separately estimate the impact of AI assistance for each group. These estimates control for year-month and agent fixed effects as in Column 5 of Table 4.2.

We find a steady and monotonic increase in returns by agent adherence: among agents in the lowest quintile, we still see a 10% gain in productivity, but for agents in the highest quintile, the estimated impact is over twice as high, close to 25%. Appendix Figure A.5 shows the results for our other four outcome measures. The positive correlation between adherence and returns holds most strongly for average handle time (Panel A) and chats per hour (Panel B), and more noisily for resolution rate (Panel C) and customer satisfaction (Panel D).

Our results are consistent with there being a treatment effect of following AI recommendations on productivity. We note, however, that this relationship could also be driven by other factors: selection (agents who choose to adhere are more productive for other reasons); or selection on gains (agents who follow recommendations are those with the greatest returns). To further explore this, we consider worker's revealed preference: do they continue to follow AI recommendations over time? If our results were driven purely by selection, we would expect workers with low adherence to continue having low adherence, since it was optimal for them to do so.

Figure 4.10 plots the evolution of AI adherence over time, for various categories of agents. Panel A begins by considering agents who differ in their initial AI compliance, which we categorize based on terciles of AI adherence in the first month of model deployment. Here, we see that compliance either stays stable or grows over time. The most initially compliant agents continue to comply at the same rates (just above 50%). Less initially compliant agents increase their compliance over time: those in the bottom tercile initially follow rec-

ommendations less than 20% of the time but, by month five, their compliance rates have increased by over 50%, to just over half of the time. Next, Panel B divides workers up by tenure at the time of AI deployment. More senior workers are initially less likely to follow AI recommendations: 30% for those with more than a year of tenure compared to 37% for those with less than three months of tenure. Over time, however, all workers increase their adherence, with more senior workers doing so faster so that the groups converge five months after deployment. In Panel C, we show the same analysis by worker skill at AI deployment. Here, we see that compliance rates are similar across skill groups, and all groups increase their compliance over time.

The results in Figure 4.10 are consistent with agents–particularly those who are initially more skeptical—coming to value AI recommendations over time. An alternative hypothesis, however, is that agents who dislike working with AI assistance exit the firm at higher rates. In Appendix Figure A.6 we repeat the analysis above, focusing on within-agent changes in adherence (that is, adherence rates residualized by agent fixed effects). Our within-agent results follow a similar pattern: all workers increase adherence over time, and these increases appear largest for workers who were initially the least compliant and workers who were the most senior. This suggests that increases in adherence over time are not driven exclusively by selection.

4.5.2 Worker Learning

A key question raised by our findings so far is whether these improvements in productivity and changes in communication patterns reflect durable changes in the human capital of workers or simply their growing reliance on AI assistance.

To study this, we examine how workers perform during periods in which they are not able to access AI-recommendations due to technical issues at the AI firm. Outages occur occasionally in our data and can last anywhere from a few minutes to a few hours. During an outage, the system fails to provide recommendations to some, but not necessarily all, workers. For example, outages may affect agents who log into their computers after the system crashes, but not agents working at the same time who had signed in earlier. They may also affect workers using one physical server but not another. Our AI firm tracks the most significant outages in order to perform technical reviews of what went wrong. We compile these system reports to identify periods in which a significant fraction of chats are impacted by outages.

Appendix Figure A.7 shows an example of such an outage, which occurred on September 10, 2020. The y-axis plots the share of post-treatment chats (e.g. those occurring after the AI system has been deployed for a given agent) for which the AI software does not provide any suggestions, aggregated to the hour level. The x-axis tracks hours in days leading up to and following the outage event (hours with fewer than 15 post-treatment chats are plotted as zeros for figure clarity). During non-outage periods, the share of chats without AI recommendations is typically 30-40%. This reflects the fact that the AI system does not generate recommendations in response to all messages, even when it is functioning properly. Because many chats are short, it is common to see chats end without the AI system intervening. On the morning of September 10th, however, we see a notable spike in the number of chats without recommendations, increasing to almost 100%. Records from

our AI firm indicate that this outage was caused by a software engineer running a load test that crashed the system.

Figure 4.11 examines the impact of access to the AI system for chats that occur during and outside these outage periods. Whereas our main event study regressions are at the worker-month level, these are at the chat level, in order to more precisely compare conversations that occurred during outage periods, versus those that did not. Panel A considers the impact of the introduction of AI assistance on chat duration (shorter is more efficient), using only post-adoption periods in which no outages are reported. Consistent with our main results, we see an immediate decline in the duration of individual chats by approximately 10% to 15%.

In Panel B, we use the same pre-treatment observations, but now restrict to post-adoption periods that are impacted by large outages. We first note that our estimates are noisy and their magnitude appears larger than for non-outage periods (15% to 25% declines in chat duration). Because AI outages are rare and not necessarily random, this may reflect differences in the types of chats that are seen during outage periods than during non-outage periods. However, focusing on the size of estimated effects over time, an interesting pattern emerges. Rather than declining immediately post-adoption and staying largely stable as we see in Panel A for non-outage periods, Panel B shows that the benefit of exposure to AI assistance increases with time during outage periods. That is, if an outage occurs one month after AI adoption, workers do not handle the chat much more quickly than their pre-adoption baseline. Yet, if an outage occurs after three months of exposure to AI recommendations, workers handle the chat faster—even though they are not receiving direct AI assistance.

In Figure 4.12, we split our main outage event studies by worker's initial AI adherence, as described in Section 4.5.1. Panel A shows that workers with high initial AI adherence see large and fast declines in chat processing times (relative to their pre-adoption baseline), even during outages. Panel B, in contrast, shows no such impact for workers who tend to deviate from AI recommendations: they see no improvement in chat times during outage periods, even after many months of AI access. These findings suggest that workers learn more by actively using AI suggestions.

Together, these results suggest that generative AI tools can help workers develop durable skills. Prior to the deployment of AI-assistance, agents only received training from managers during brief weekly coaching sessions. During these sessions, managers would go through several conversations from the past week and advise the worker on how they might have handled certain conversations better. However, by necessity, managers can only provide feedback on a small fraction of the conversations that an agent conducts. Moreover, because managers are often pressed for time and may lack training, they may simply point out weak metrics ("you need to lower your handle time") rather than identifying strategies for how an agent could better approach a problem ("you need to ask more questions at the beginning to diagnose the issue better.") This type of coaching is ineffective and can be counterproductive for employee engagement (Berg et al., 2018). In contrast, AI assistance provides workers with specific, actionable suggestions in real time. Our findings suggest that this can play a useful role in supplementing existing on-the-job training programs.

4.5.3 Conversational Change

Lastly, we consider how access to AI assistance influences how workers communicate. To capture an overall sense of the content of conversations, we begin by creating textual embeddings of agent-customer conversations. Textual embeddings take a given body of text and transform it into a high-dimensional vector that represents its "coordinates" in linguistic space. Two pieces of text will have more similar coordinates if they share a common meaning or style. The specific embedding given to a body of text will depend on the embedding model that is used. We form our text embeddings using all-MiniLM-L6-v2, an LLM that is specifically intended to capture and cluster semantic information to assess similarity across text (Hugging Face, 2023). Once we create an embedding for each conversation, we can compare the similarity of conversations by looking at the cosine similarities of their associated vectors; this common approach yields a score of 0 if two pieces of text are semantically orthogonal and a score of 1 if they have the same meaning (Koroteev, 2021). For context, the sentences "Can you help me with logging in?" and "Why is my login not working?" have a cosine similarity of 0.68 in our model.

Using this approach, we first show that AI assistance changes the content of what agents write to customer, rather than just typing the same things faster. Second, we explore how these patterns differ for high- and low-skill workers. We are particularly interested in understanding whether AI models can disseminate the behaviors of high performers. If this is the case, then we would expect AI assistance to lead lower-performing agents to write more like high-performers.

Within-worker changes in communication

We begin by examining how an agent's communication evolves over time, before and after access to AI assistance. We begin by examining treated workers and comparing the similarity of their chats in each given event-time week to their chats from the month before AI deployment (week -4 to week -1). We exclude messages from the customer and focus only on agent-generated language. Panel A of Figure 4.13 plots the cosine similarity associated with these comparisons, We find that textual similarity to the pre-AI window is stable in the weeks leading up to the AI roll-out and drops immediately following AI deployment. That is, conversations 12 to 5 weeks before deployment are quite similar to conversations 4 to 1 week before, but conversations 0 to 12 weeks after are all less similar.

This drop in similarity is broadly inconsistent with the idea that AI assistance merely leads workers to type the same things but faster. If that were the case, we would expect call handle times to drop, but textual similarity to remain constant.

Next, Panel B of Figure 4.13 compares the magnitude of this pre- versus post-deployment change in textual content varies by pre-AI worker skill. We find that lower-skill agents (those in the bottom quintile of the pre-AI skill distribution) experience greater textual change after AI adoption, relative to top performers (those in the top quintile). Our results here control for firm-year-month fixed effects, which can account for seasonal changes in topics such as tax or payroll cycles, or new product rollouts. We also control for agent tenure fixed effects, which can account for the possibility that younger workers' language may evolve more quickly independent of access to the AI model. Although we cannot control directly

for the chat topic, we note that chat topics are randomly assigned to agents, so we would not expect differences in topics to vary systematically by agent skill. We interpret these results as providing suggestive evidence that AI deployment shifts the communication patterns of low-skill workers more than high-skill workers.

Across worker comparisons

Figure 4.14 considers whether individual level changes in communication lead low- and high-skill workers to sound more alike. To examine this, we plot the cosine similarity between high- and low-skill agents at specific moments in calendar time, separately for workers with (blue dots) and without (red diamonds) access to AI assistance. Among agents without AI access, we define high- and low-skill agents as those who are in the top or bottom quintile of our skill index for that month. Among agents with AI access, we define high- and low-skill agents based on whether they are in the top or bottom quintile of skill at the time of AI deployment.

Focusing on the blue dots, we see that the average textual similarity between high- and low-productivity workers is 0.55 among workers who do not have access to AI assistance. This figure is lower than our average within-person text similarity, which makes sense given that within-person changes are likely to be smaller than across-person differences. We see, moreover, that this textual similarity is stable over time, indicating that high- and low-skill workers do not appear to be trending differently in the absence of AI assistance.

Turning to the red diamonds, we see that, post-AI adoption, high- and low-skilled workers begin to use language that is more similar. The magnitude of this change—moving from 0.55 similarity to 0.61 similarity—may appear small, but given that the average within-person similarity for high-skill workers is around 0.67, this result suggests that AI assistance is associated with a substantial narrowing of language gaps.

Together, the patterns in Figures 4.13 and 4.14 suggest that low-skill workers are converging toward high-skill workers, rather than the opposite. This finding is consistent with AI models disseminating the behaviors of high-skilled workers to lower-skilled workers. In such a scenario, we would expect low-skill workers to change their communication patterns more following AI-deployment. Top performers, meanwhile, would change less because the AI model is more likely to suggest language they already use.

4.6 Effects on the Experience of Work

Qualitative studies suggest that working conditions for contact center agents can be unpleasant. The repetitive nature of the job, coupled with regular exposure to challenging and emotionally charged conversations, can contribute to burnout and high turnover rates. Additionally, contact center work for US-based businesses is frequently outsourced to lower-income countries such as India and the Philippines, meaning that agents often work difficult hours and may face cultural barriers or judgements when speaking with customers.

Increases in worker productivity may not neccesarily lead workers to be happier with their jobs, especially if workers feel pressured to work faster and faster. In this section, we examine the impact of generative AI on one key aspect of the workplace experience: how

agents are treated by customers, as measured by customer sentiment and requests to speak with a manager. We also examine the impact of AI assistance on worker turnover as an overall indicator of worker satisfaction.

4.6.1 Customer Sentiment

Customers often vent their frustrations on anonymous service agents and, in our data, we see regular instances of swearing, verbal abuse, and "yelling" (typing in all caps). Service workers are called upon to absorb such customer frustrations while limiting one's own emotional reaction (Hochschild, 2019). The stress associated with this type of emotional labor is often cited as a key cause of burnout and attrition among customer service workers (Lee, 2015).

Access to AI-assistance may impact how customers treat agents, but the direction and magnitude of these impacts are ambiguous. AI assistance may improve the tenor of conversations by helping agents set customer expectations or resolve their problems more quickly. Alternatively, customers may become more frustrated if AI-suggested language feels "corporate" or insincere.

To assess this, we attempt to capture the affective nature of both agent and customer text, using sentiment analysis (Mejova, 2009). For this analysis, we use SiEBERT, an LLM that is fine-tuned for sentiment analysis using a variety of datasets, including product reviews and tweets (Hartmann et al., 2023). Sentiment is measured on a scale from -1 to 1, where -1 indicates negative sentiment and 1 indicates positive. In a given conversation, we compute separate sentiment scores for both agent and customer text. We then aggregate these chat-level variables into a measure of average agent sentiment and average customer sentiment for each agent-year-month.

Panel A of Figure 4.15 shows the distribution of customer sentiment scores. On average, customer sentiments in our data are mildly positive and normally distributed around a mean of 0.14, except for a mass of very positive and very negative scores. Panel B shows the distribution of sentiments associated with agents: agents are unfailingly positive, with a mean sentiment score of 0.89. This reflects the fact that agents are trained to be extremely polite and friendly, even prior to AI access.

Panels C and D consider how sentiment scores respond following the roll-out of AI assistance. In Panel C, we see an immediate and persistent improvement in customer sentiment. This effect is economically large: according to Column 1 of Table 4.4, access to AI improves the mean customer sentiments (averaged over an agent-month) by 0.18 points, equivalent to half of a standard deviation. In Panel D, we see no detectable effect for agent sentiment, which is already very high at baseline. Column 2 of Table 4.4 indicates that agent sentiments increase by only 0.02 points or about 1% of a standard deviation.

Focusing on customer sentiment, Appendix Figure A.8 examines whether access to AI has different impacts for across agents. We find that access to AI assistance significantly improves how customers treat agents of all skill and experience levels, with the largest effects for agents in the lower to lower-middle range of both the skill and tenure distributions. Consistent with our productivity results, the highest-performing and most-experienced agents see the smallest benefits of AI access. These results suggest that AI recommendations, which were explicitly designed to prioritize more empathetic responses, may improve agents' demonstrated social skills and have a positive emotional impact on customers.

4.6.2 Customer Confidence and Managerial Escalation

Changes in individual worker-level productivity may have broader implications for organizational workflows (Athey and Stern, 1998*b*; Athey et al., 1994; Garicano, 2000). In most customer service settings, front-line agents attempt to resolve customer problems but can seek the help of supervisors when they are unsure of how to proceed. Customers, knowing this, will sometimes attempt to escalate a conversation by asking to speak to a manager. This type of request generally occurs when the customer feels that the current agent is not equipped to address their problem or becomes frustrated.

In Figure 4.16, consider the impact of access to AI-assistance on the frequency of chat escalation. The outcome variable we focus on is the share of an agent's chats in which a customer requests to speak to a manager or supervisor, aggregated to the year-month level. We focus on requests for escalation rather than actual escalations both because we lack data on actual escalations and because requests are a better measure of customer confidence in an agent's competence or authority. Following the introduction of AI assistance, we see a gradual decline in requests for escalation. Relative to a baseline rate of approximately 6 percentage points, these coefficients suggest that AI assistance generates an almost 25% decline in customer requests to speak to a manager. In Appendix Figure A.9, we consider how these patterns change by the skill and experience of the worker. Consistent with our other results, we find that requests for escalation are disproportionately reduced for agents who were less skilled or less experienced at the time of AI adoption.

4.6.3 Attrition

The adoption of generative AI tools can have a variety of impacts on workers: their productivity, the amount of stress they encounter on the job, and how they are perceived by customers, to name a few. While we cannot observe all these factors, we can look at turnover patterns to provide one overarching measure of how workers are impacted by AI technology at work.

For this analysis, we compare attrition rates among treated agents to those of untreated agents with the same tenure. We drop observations for treated agents before treatment because they do not experience attrition by construction (they must survive to be treated in the future). Our analysis also controls for location and time fixed effects. Figure 4.17 plots the impact of AI access on attrition: Panel A considers how this varies by agent tenure while Panel B considers heterogeneity by agent skill. Consistent with our findings so far, Panel A shows that access to AI assistance is associated with the strongest reductions in attrition among newer agents, those with less than 6 months of experience. The magnitude of this coefficient, around 10 percentage points, translates into a 40% decrease relative to a baseline attrition rate in this group of 25%. In Panel B, we examine attrition by worker skill. Here, we find a significant decrease in attrition for all skill groups, but no systematic gradient.

Finally, we note that these results should be taken with more caution relative to our main results because we are unable to include agent fixed effects to control for unobservable differences between agents with and without access to AI assistance. This is because attrition can only occur once for any given individual. Our results may overstate the impact of AI access on attrition if, for example, access to the AI tool is more likely to be given to agents

whom the firm believes are more likely to stay.

4.7 Conclusion

Advancements in AI technologies open up a broad set of economic possibilities. Our paper provides the first empirical evidence of the effects of a generative AI tool in a real-world workplace. In our setting, we find that access to AI-generated recommendations increases worker productivity, improves customer sentiment, and is associated with reductions in employee turnover.

We hypothesize that part of the effect we document is driven by the AI system's ability to embody the best practices of high-skill workers in our firm and make it accessible to other workers. These practices may have previously been difficult to disseminate because they involve tacit knowledge. Consistent with this, we see that AI assistance leads to substantial improvements in problem resolution and customer satisfaction for newer- and less-skilled workers but does not help the highest-skilled or most-experienced workers on these measures. Furthermore, agents who have used the system perform somewhat better even when the system is unexpectedly disabled. Analyzing the text of agent conversations, we find suggestive evidence that AI recommendations lead low-skill workers to communicate more like high-skill workers.

Our findings, and their limitations, point to a variety of directions for future research.

Most importantly, our results do not capture the potential longer-term impacts of generative AI on skill demand, job design, wages, or customer demand. It is unclear, for example, whether improvements in customer service productivity will lead to more or less demand for customer service workers. If the demand for customer support is inelastic, then generative AI tools may reduce demand and wages in this sector in the long run. Alternatively, better product support could lead customers to seek out representatives for a wider range of questions; this, in turn, could increase demand for workers or give them new responsibilities, such as collecting customer feedback for the product development team (Berg et al., 2018; Korinek, 2022).

Our findings also raise questions about the nature of worker productivity. Traditionally, a support agent's productivity refers to their ability to help the customers they come in contact with. Yet, in a setting where customer service conversations are fed into training datasets, a worker's productivity also includes their ability to provide ML models with examples of successful behaviors that can be shared with others. In our setting, top performers contribute many of the examples used to train the AI system we study, but they see relatively few improvements in their own productivity as a result. Under our data firm's current pay practices, these workers may even see a reduction in their pay because bonuses are calculated relative to other agents' performance. Our results therefore raise questions about how workers, particularly top performers, should be compensated for the data that they provide to AI systems.

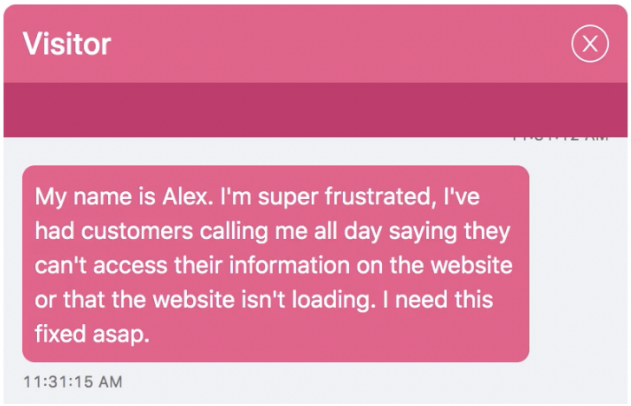
Finally, as a potential general-purpose technology, generative AI can and will be deployed in a variety of ways, and the effects we find may not generalize across all firms and production processes (Eloundou et al., 2023). For example, our setting has a relatively stable product and a set of technical support questions. In areas where the product or environment is

changing rapidly, the relative value of AI recommendations may be different: they may be better able to synthesize changing best practices, or they may actually impede learning by promoting outdated practices observed in historical training data.

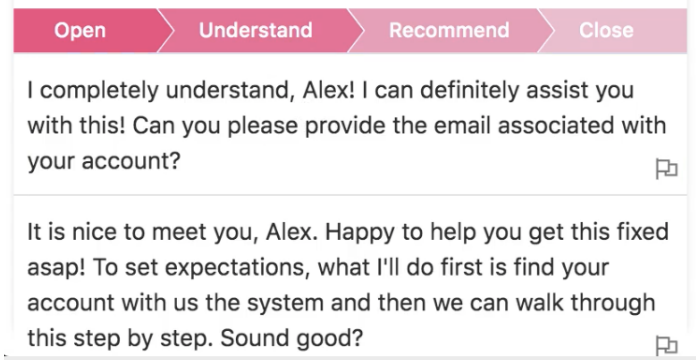
Given the early stage of generative AI, these and other questions deserve further scrutiny.

FIGURE 4.1: SAMPLE AI OUTPUT

A. SAMPLE CUSTOMER ISSUE

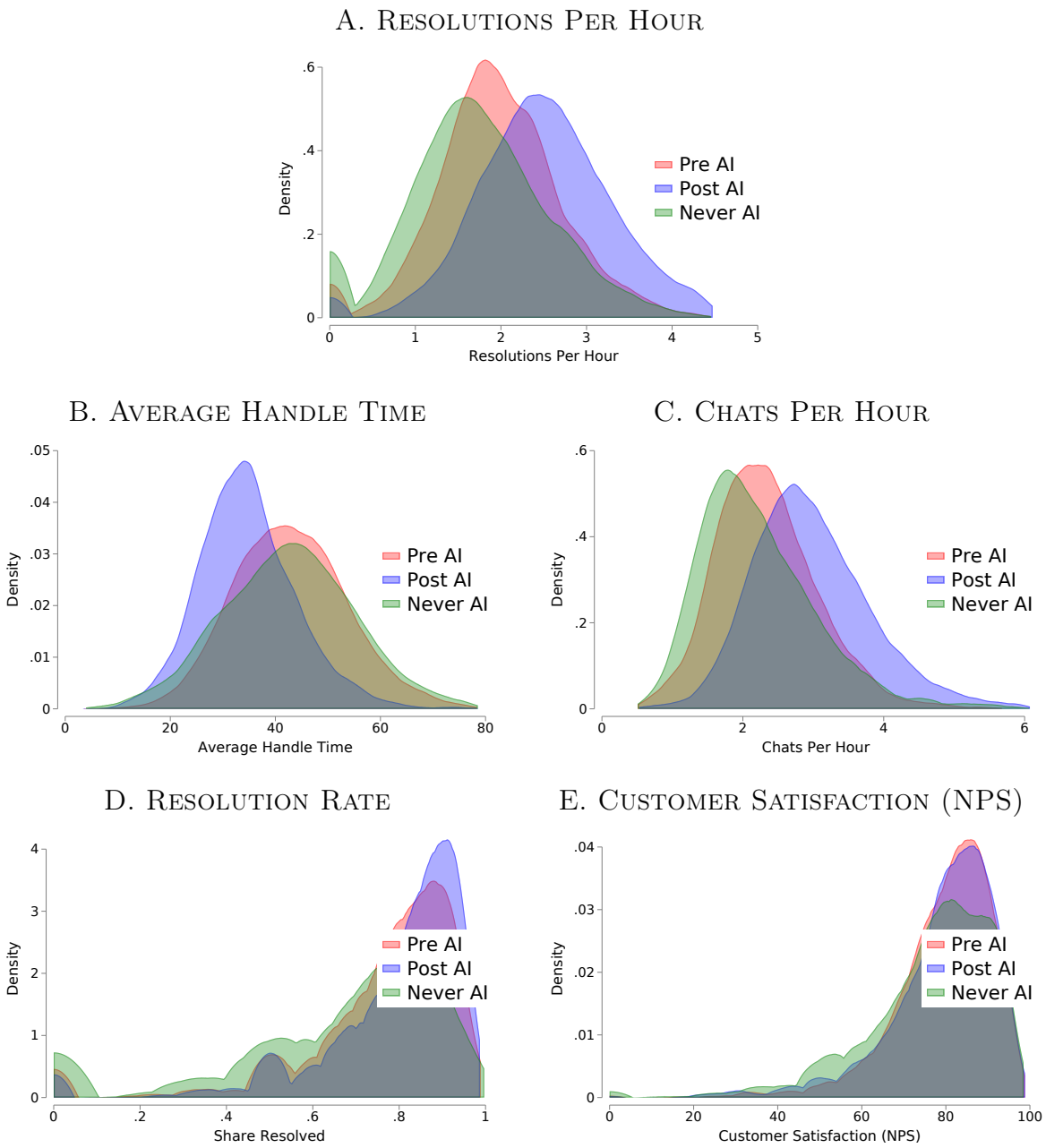


B. SAMPLE AI-GENERATED SUGGESTED RESPONSE



NOTES: This figure shows sample suggestions of output generated by the AI model. The suggested responses are only visible to the agent. Workers can choose to ignore, accept or somewhat incorporate the AI suggestions into their response to the customer.

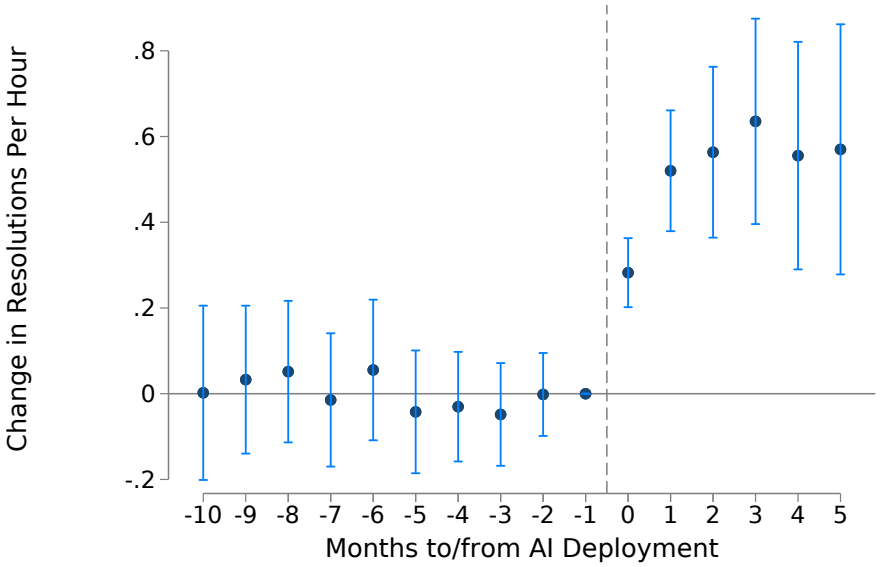
FIGURE 4.2: RAW PRODUCTIVITY DISTRIBUTIONS, BY AI TREATMENT



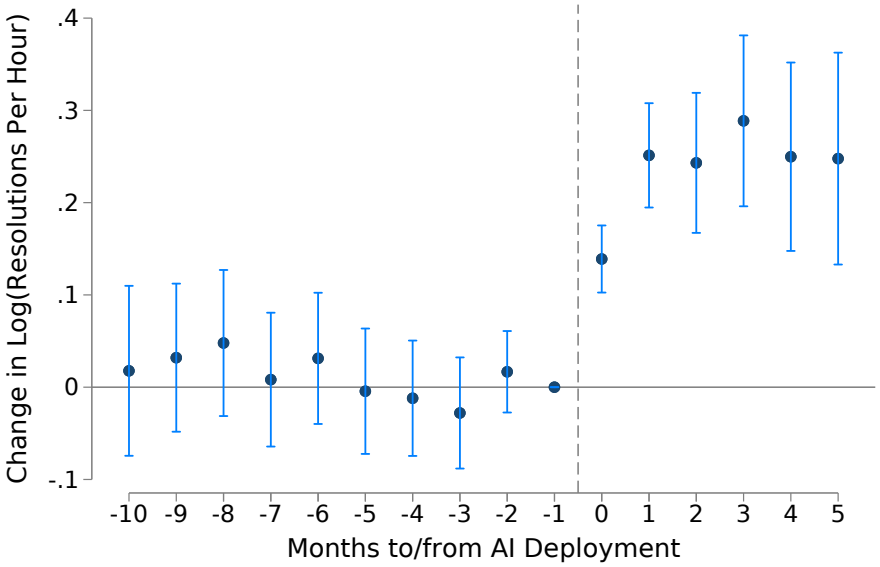
NOTES: This figure shows the distribution various outcome measures. We split this sample into agent-month observations for agents who eventually receive access to the AI system before deployment ("Pre AI"), after deployment ("Post AI"), and for agent-months associated with agents who never receive access ("Never AI"). Our primary productivity measure is "resolutions per hour," the number of customer issues the agent is able to successfully resolve per hour. We also provide descriptives for "average handle time," the average length of time an agent takes to finish a chat; "chats per hour," the number of chats completed per hour incorporating multitasking; "resolution rate," the share of conversations that the agent is able to resolve successfully; and "net promoter score" (NPS), which are calculated by randomly surveying customers and calculating the percentage of customers who would recommend an agent minus the share who would not.

FIGURE 4.3: EVENT STUDIES, RESOLUTIONS PER HOUR

A. RESOLUTIONS PER HOUR

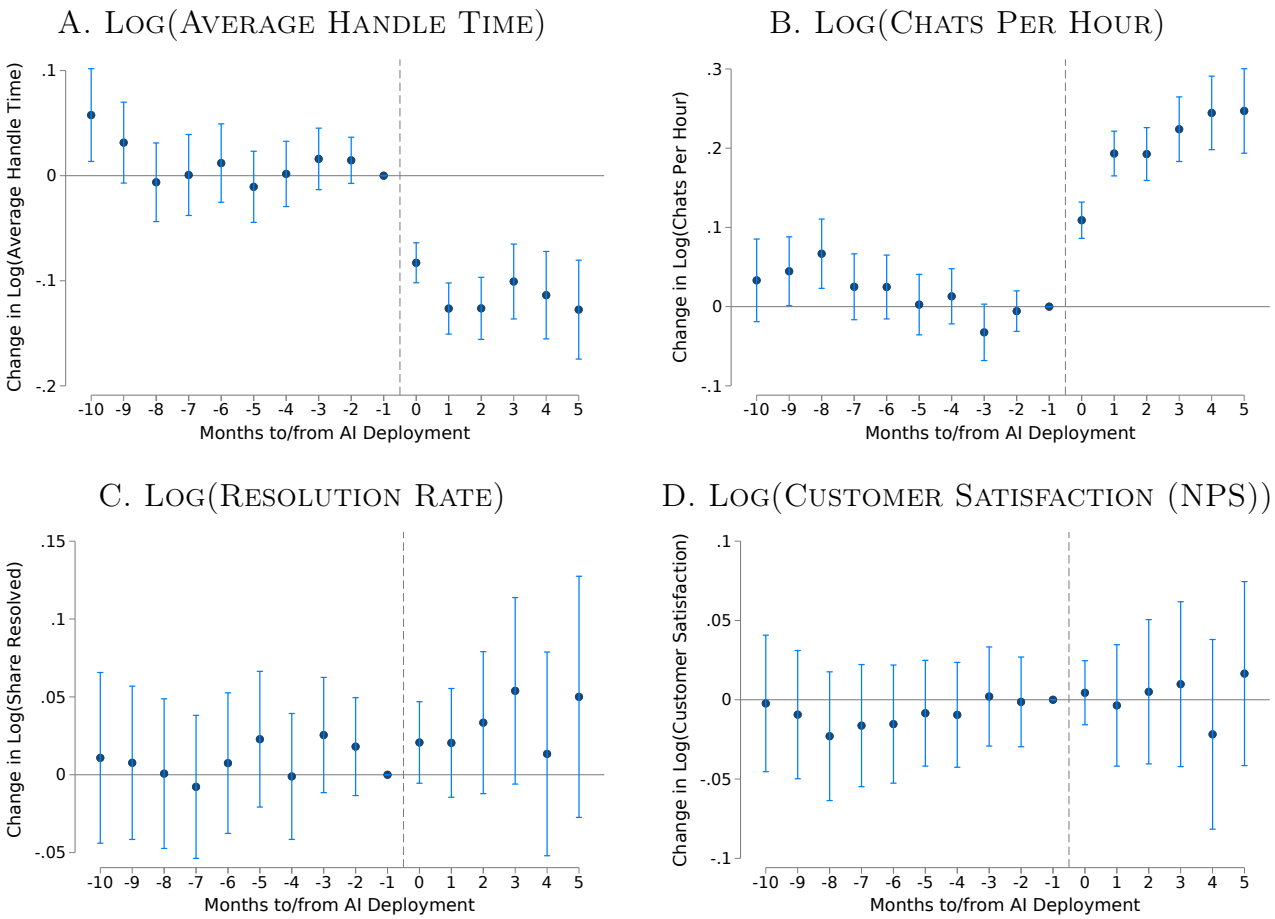


B. LOG(RESOLUTIONS PER HOUR)



NOTES: These figures plot the coefficients and 95% confidence intervals from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator. See text for additional details. Panel A plots the resolutions per hour and Panel B plots the natural log of the measure. All specifications include agent and chat year-month, location, agent tenure and company fixed effects. Robust standard errors are clustered at the agent level.

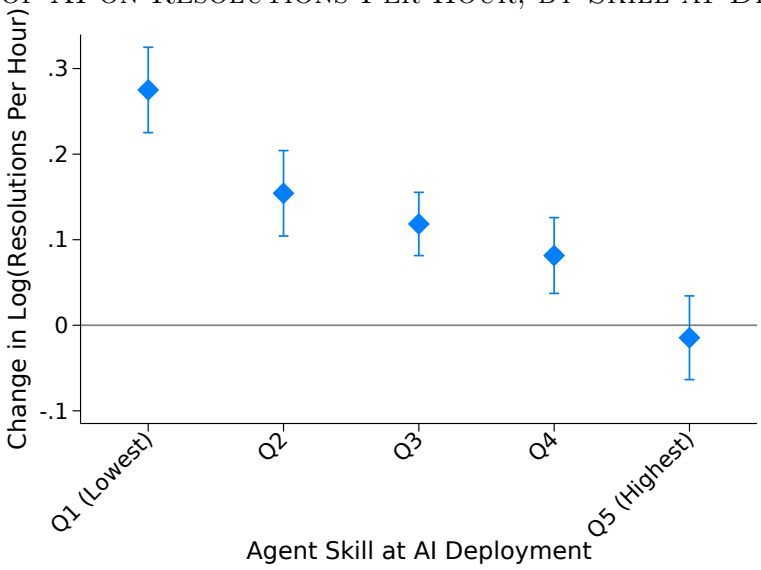
Figure 4.4: Event Studies, Additional Outcomes



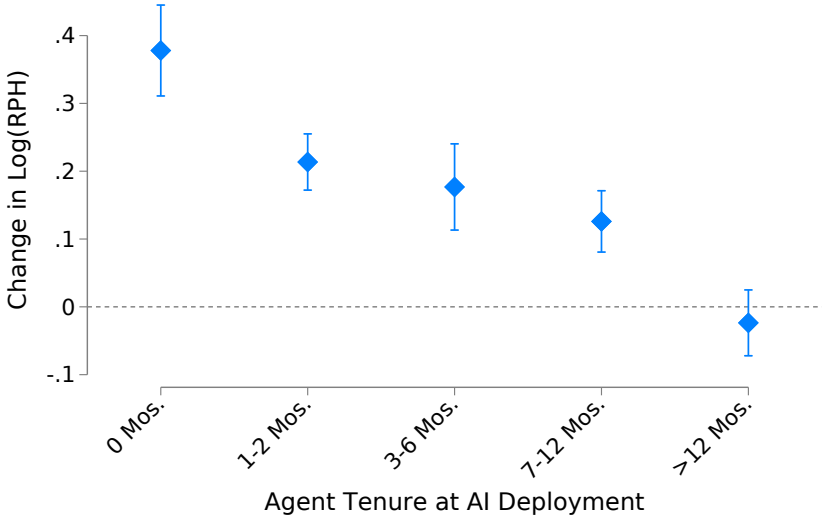
Notes: These figures plot the coefficients and 95% confidence intervals from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator. See text for additional details. Panel A plots the average handle time or the average duration of each technical support chat. Panel B plots the number of chats an agent completes per hour, incorporating multitasking. Panel C plots the resolution rate, the share of chats successfully resolved, and Panel D plots net promoter score, which is an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, agent tenure and company fixed effects. Robust standard errors are clustered at the agent level.

FIGURE 4.5: HETEROGENEITY OF AI IMPACT, BY SKILL AND TENURE

A. IMPACT OF AI ON RESOLUTIONS PER HOUR, BY SKILL AT DEPLOYMENT

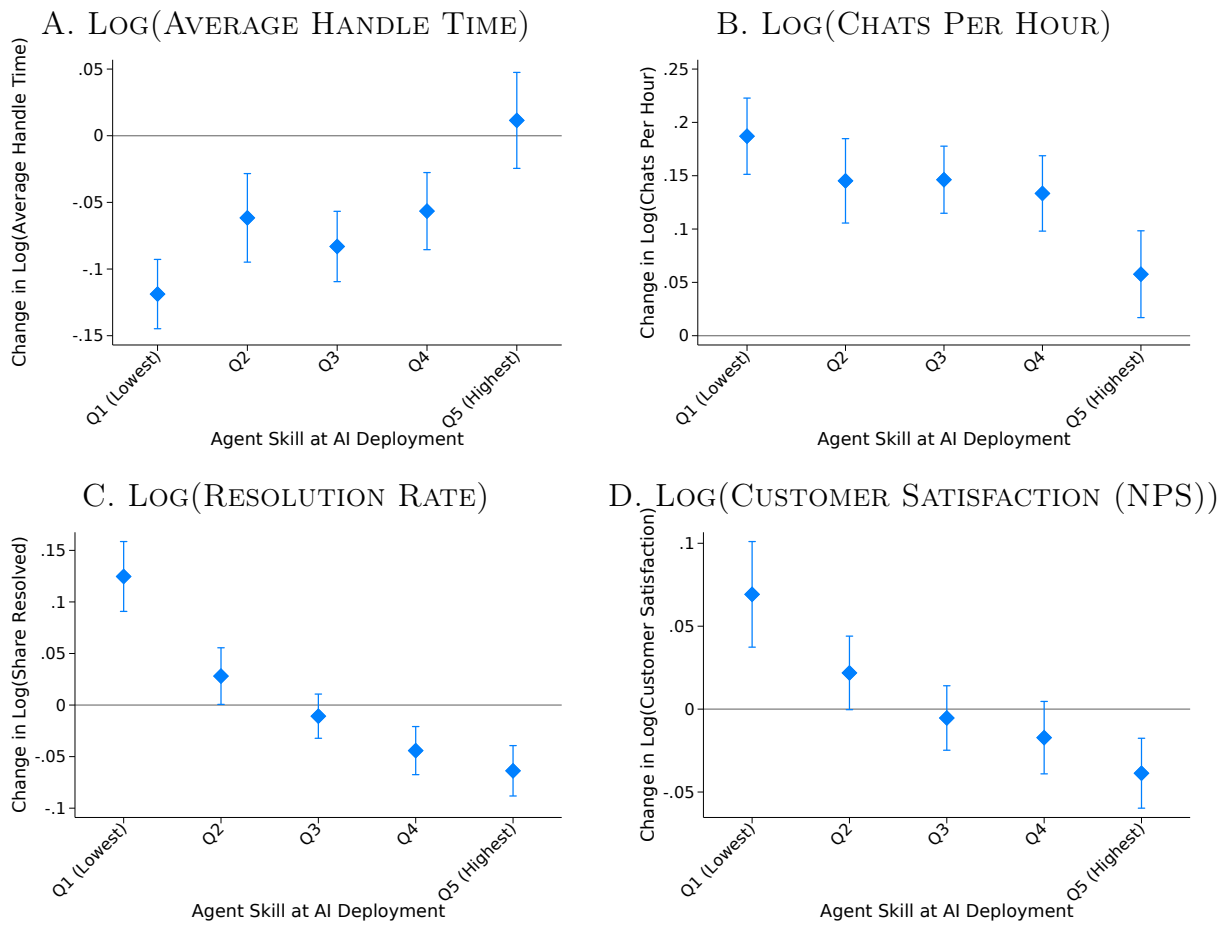


B. IMPACT OF AI ON RESOLUTIONS PER HOUR, BY TENURE AT DEPLOYMENT



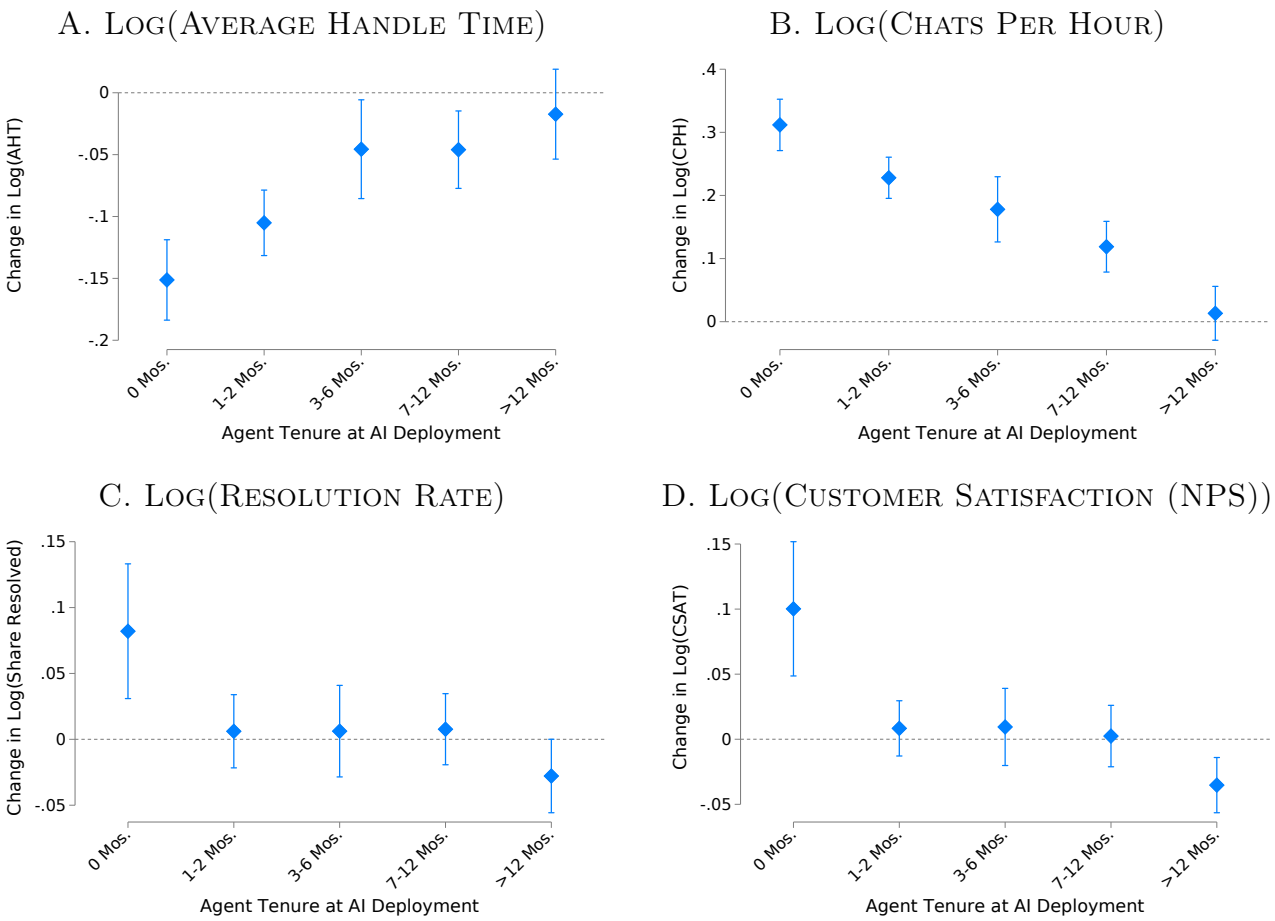
NOTES: These figures plot the impacts of AI model deployment on log(resolutions per hour) for different groups of agents. Agent skill is calculated as the agent's trailing three month average of performance on average handle time, call resolution, and customer satisfaction, the three metrics our firm uses to assess agent performance. Within each month and company, agents are grouped into quintiles, with the most productive agents in quintile 5 and the least productive in quintile 1. Pre-AI worker tenure is the number of months an agent has been employed when they receive access to AI recommendations. All specifications include agent and chat year-month, location, and company fixed effects and standard errors are clustered at the agent level. Panel A includes controls for agent tenure at deployment and Panel B includes controls for agent skill at deployment.

Figure 4.6: Heterogeneity of AI Impact by pre-AI Worker Skill and Controlling for Tenure, Additional Outcomes



Notes: These figures plot the impacts of AI model deployment on four measures of productivity and performance, by pre-deployment worker skill. Agent skill is calculated as the agent's trailing three month average of performance on average handle time, call resolution, and customer satisfaction, the three metrics our firm uses for agent performance. Within each month and company, agents are grouped into quintiles, with the most productive agents within each firm in quintile 5 and the least productive in quintile 1. Panel A plots the average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of chats an agent can handle per hour. Panel C plots the resolution rate, and Panel D plots net promoter score, an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, and company fixed effects, controls for agent tenure and standard errors are clustered at the agent level.

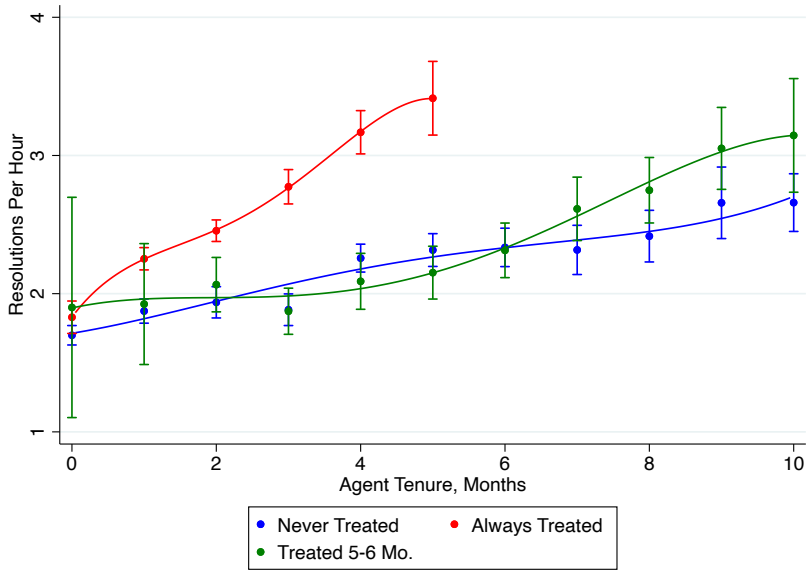
FIGURE 4.7: HETEROGENEITY OF AI IMPACT BY PRE-AI WORKER TENURE CONTROLLING FOR SKILL, ADDITIONAL OUTCOMES



NOTES: These figures plot the impacts of AI model deployment on measures of productivity and performance by pre-AI worker tenure, defined as the number of months an agent has been employed when they receive access to the AI model. Panel A plots the average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of chats an agent can handle per hour. Panel C plots the resolution rate, and Panel D plots net promoter score, an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, and company fixed effects, controls for agent skill at deployment and standard errors are clustered at the agent level.

FIGURE 4.8: EXPERIENCE CURVES BY DEPLOYMENT COHORT

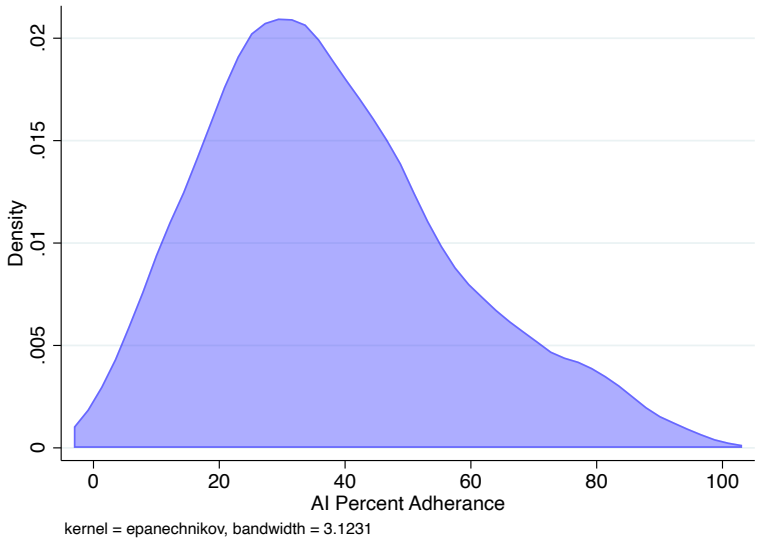
RESOLUTIONS PER HOUR, BY AGENT TENURE



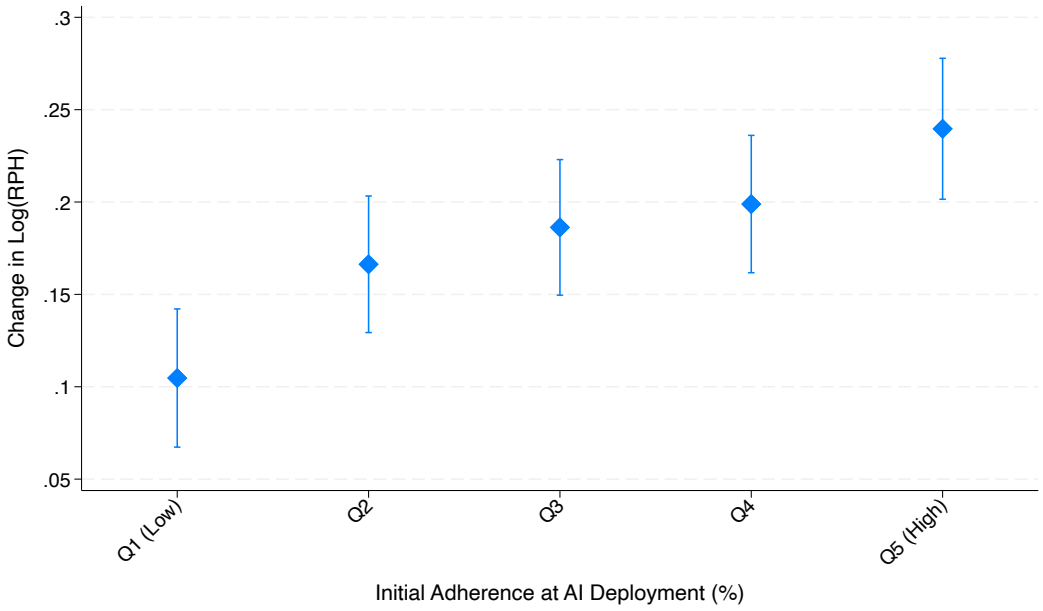
NOTES: This figure plot the relationship between productivity and job tenure. The red line plots the performance of always-treated agents, those who have access to AI assistance from their first month on the job. The blue line plots agents who are never treated. The green line plots agents who spend their first four months of work without the AI assistance, and gain access to the AI model during their fifth month on the job. 95% confidence intervals are shown.

FIGURE 4.9: HETEROGENEITY OF AI IMPACT, BY AI ADHERENCE

A. DISTRIBUTION OF AI ADHERENCE



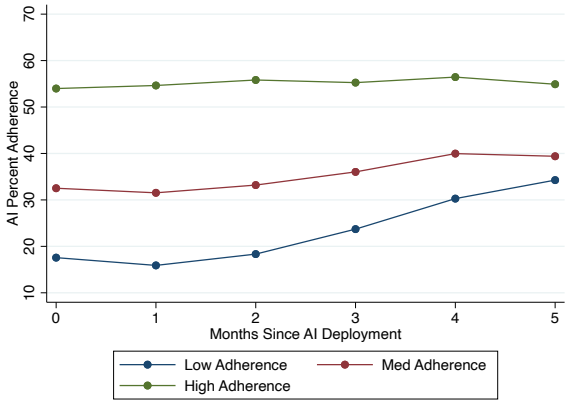
B. IMPACT OF AI ON RESOLUTIONS PER HOUR, BY INITIAL ADHERENCE



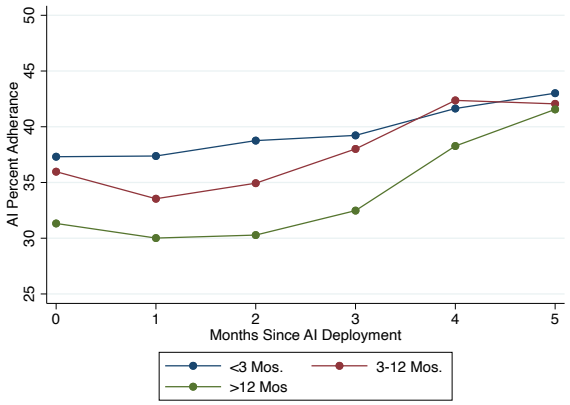
NOTES: Panel A plots the distribution of AI adherence, averaged at the agent-month level, weighted by the log of the number of AI recommendations for that agent-month. Panel B shows the impact of AI assistance on resolutions by hour, by agents grouped by their initial adherence, defined as the share of AI recommendations they followed in the first month of treatment.

Figure 4.10: AI Adherence over Time

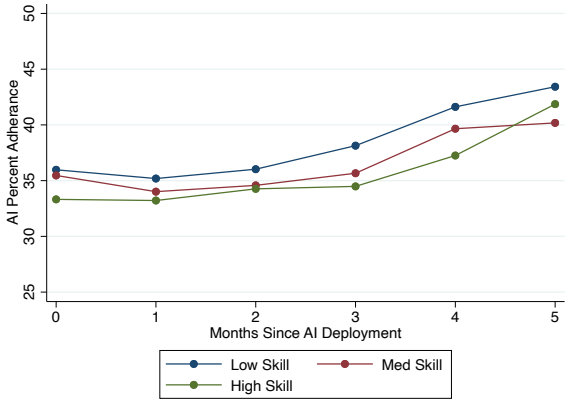
A. By Adherence at AI Model Deployment



B. By Agent Tenure at AI Model Deployment



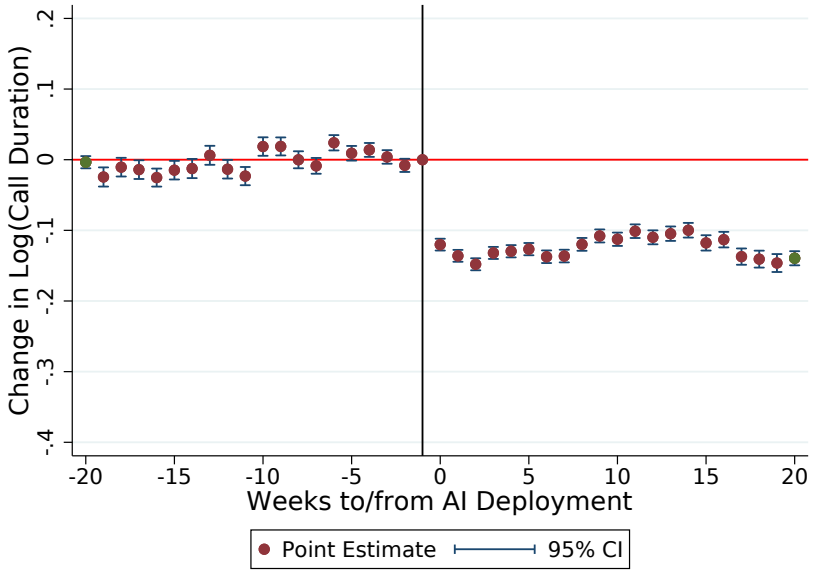
C. By Agent Skill at AI Model Deployment



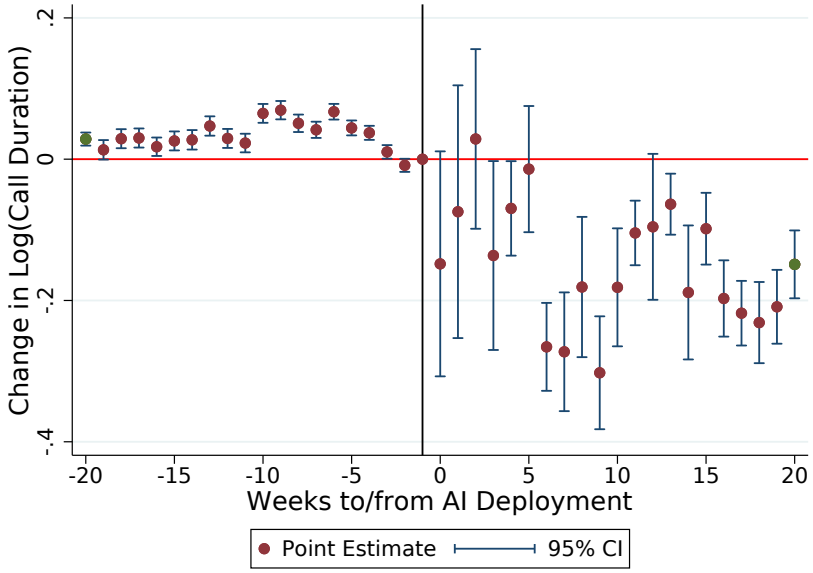
Notes: This figure plots the share of AI suggestions followed by agents as a function of the number of months each agent has had access to the AI. In Panel A, we divide agents into terciles based on their adherence to AI suggestions in the first month. In Panel B, we divide agents into groups based on their tenure at the firm at the time of AI model deployment. In Panel C, we divide workers into terciles with our skill index.

Figure 4.11: Productivity during AI System Outages

A. Post-treatment Non-Outage Periods



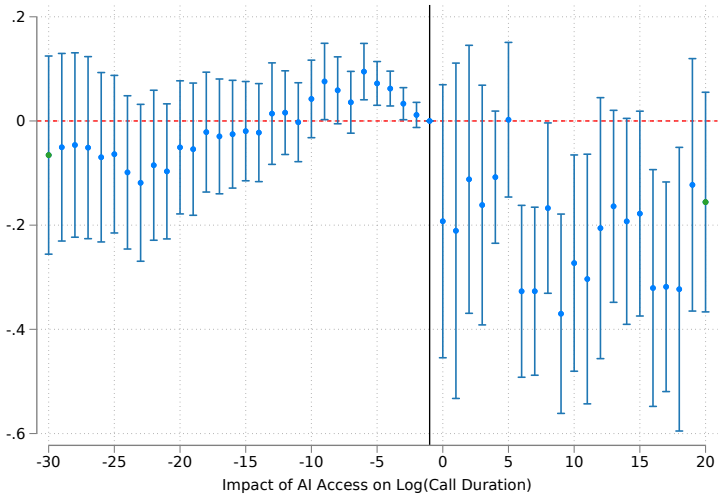
B. Post-treatment Outage Periods



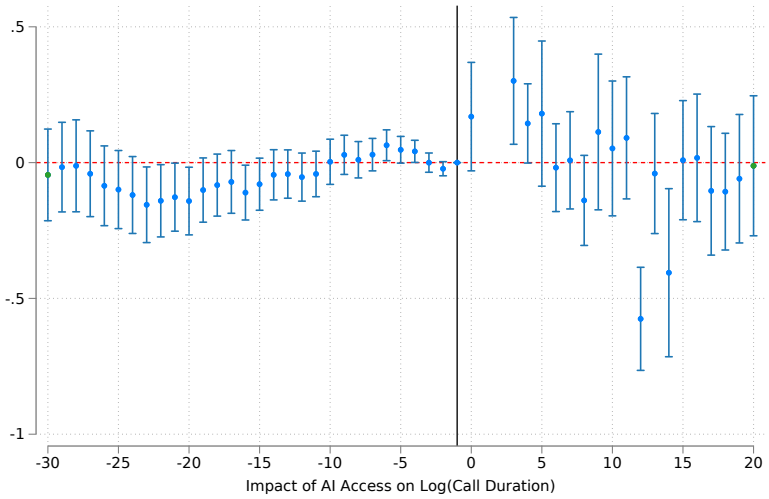
Notes: This figure plots event studies of the impact of the roll out of AI-assistance on chat times at the individual chat level. Panel A restricts to post-treatment chats that do not occur during any period where there is a system outage. Panel B restricts to post-treatment chats that only occur during a large system outage. Standard errors are clustered at the agent level.

FIGURE 4.12: PRODUCTIVITY DURING AI SYSTEM OUTAGES, BY INITIAL AI ADHERENCE

A. HIGH INITIAL ADHERENCE



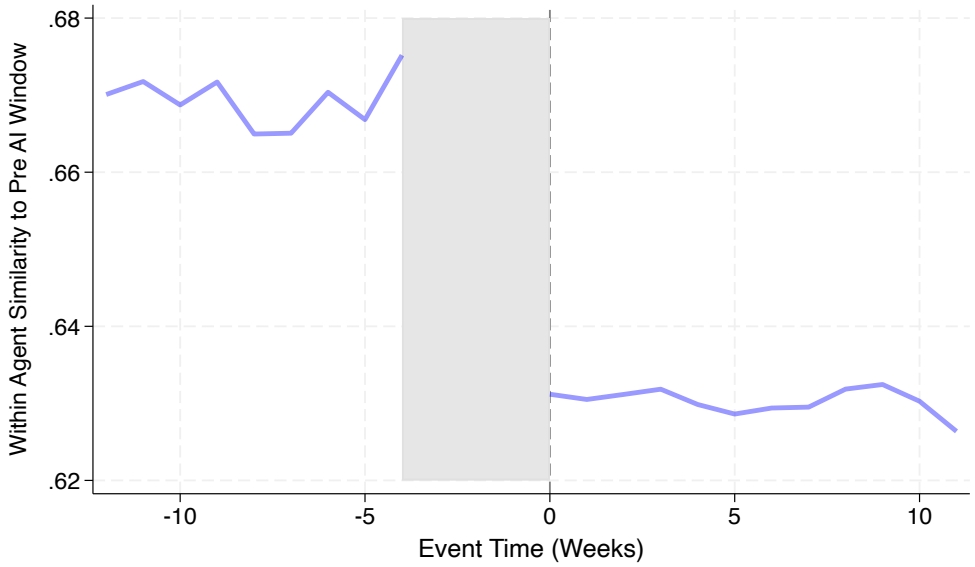
B. LOW INITIAL ADHERENCE



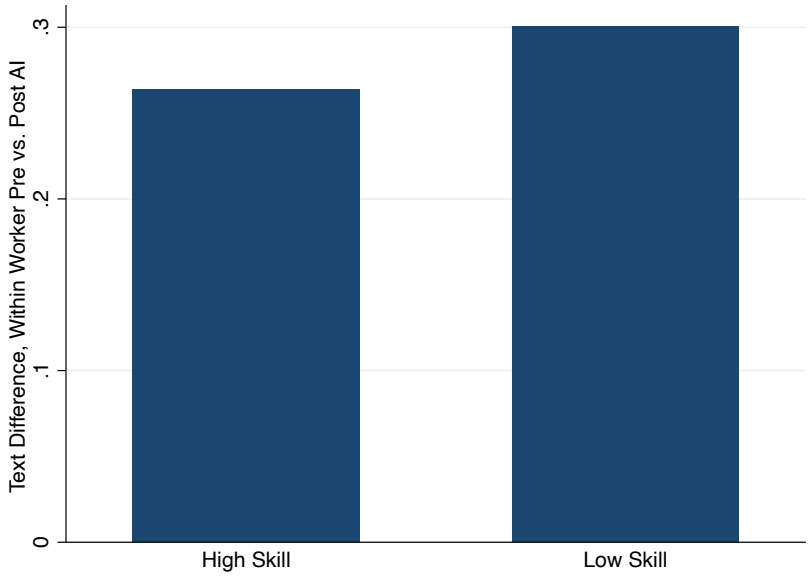
NOTES: This figure plots event studies for the impact of AI system rollout of chat duration. All post-adoption data is restricted to periods in which there is a major outage. Panel A restricts to only chats generated by ever-treated agents who with high initial AI adherence (top tercile) while Panel B restricts to agents with low initial adherence (bottom tercile). Agents who are never treated are excluded from this analysis.

FIGURE 4.13: WITHIN AGENT TEXTUAL ANALYSIS

A. WITHIN-PERSON TEXTUAL SIMILARITY TO MONTH PRIOR TO AI

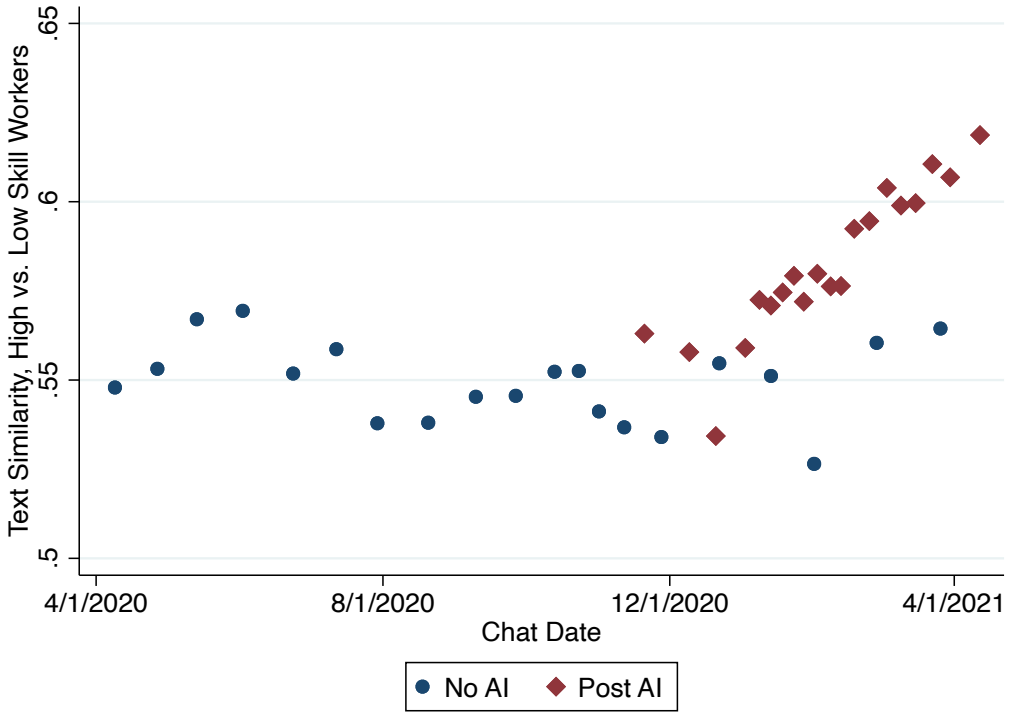


B. WITHIN-PERSON TEXTUAL CHANGE, LOW VS. HIGH SKILL



NOTES: Panel A plots the average similarity between an agent's chats each week and a comparison group of their conversations in the month prior to AI deployment. To avoid comparing conversations to themselves, we exclude calculating the similarity in the month prior to deployment. Panel B plots the average difference between an agent's pre-AI corpus of chat messages and that same agent's post-AI corpus, controlling for year-month and agent tenure. The first bar represents the average pre-post text difference for agents in the highest quintile of pre-AI skill, as measured by a weighted index of their chats per hour, resolution rate, and customer satisfaction score. The low-skill bar represents the same type of pre-post text difference among the lowest skill quintile. Agent skill, or relative productivity, is defined at the time of treatment.

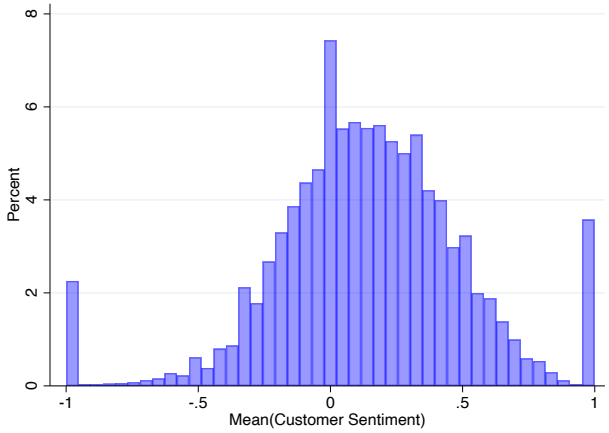
FIGURE 4.14: TEXT SIMILARITY BETWEEN LOW-SKILL AND HIGH-SKILL WORKERS, PRE AND POST AI



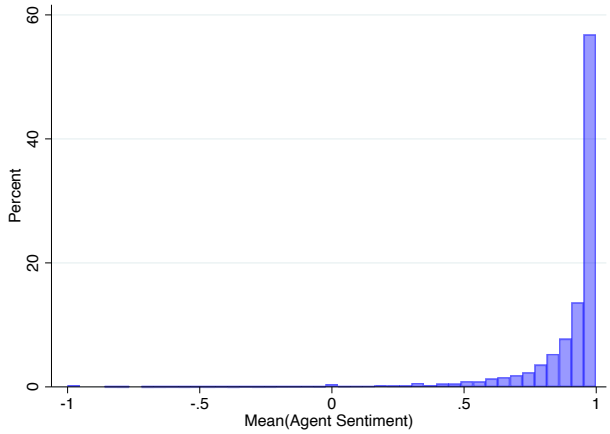
NOTES: This figure plots the average text similarity between the top and bottom quintile of agents. The blue line plots the similarity for never treated or pre-treatment agents, the red line plots the similarity for agents with access to the AI model. For agents in the treatment group, we define agent skill prior to AI model deployment. Our analysis includes controls for agent tenure.

FIGURE 4.15: CONVERSATION SENTIMENT

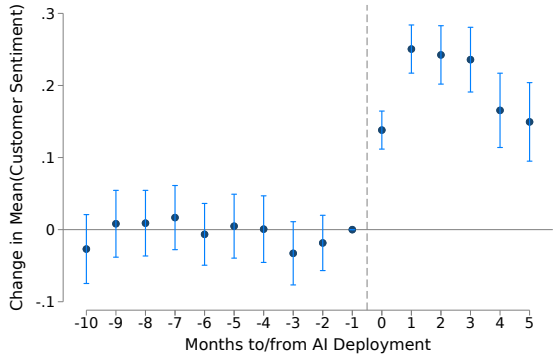
A. CUSTOMER SENTIMENT, HISTOGRAM



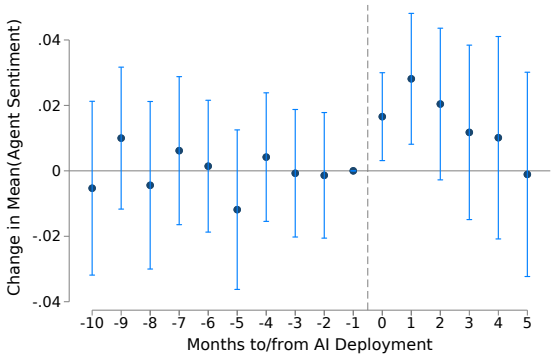
B. AGENT SENTIMENT



C. CUSTOMER SENTIMENT, EVENT STUDY

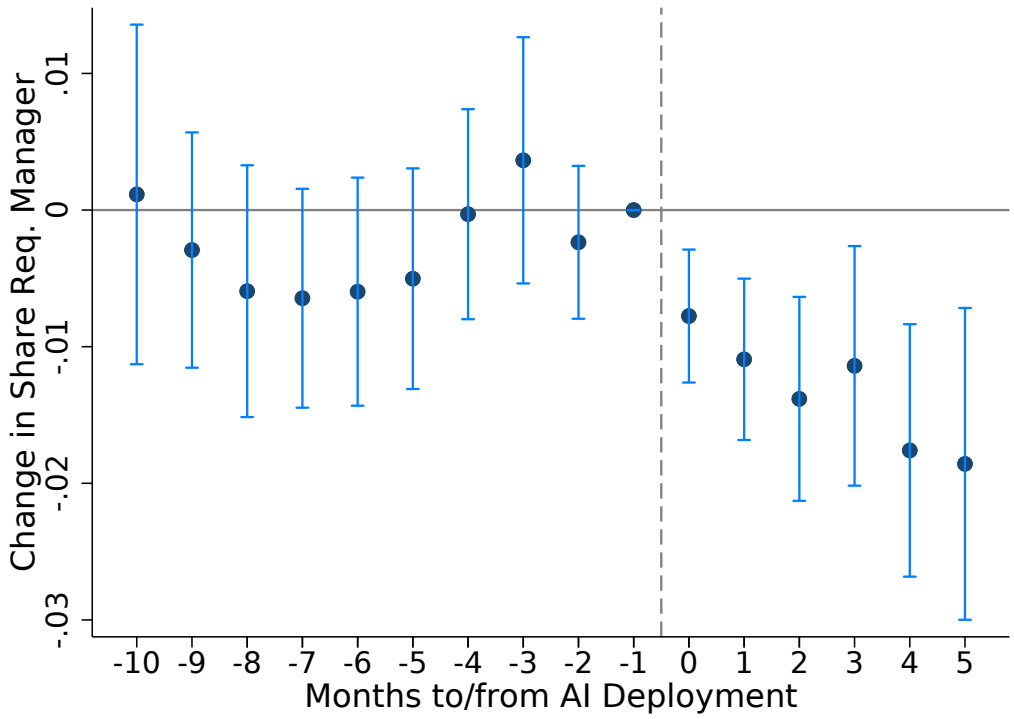


D. AGENT SENTIMENT, EVENT STUDY



NOTES: Each panel of this figure plots the impact of AI model deployment on conversational sentiment. Panel A shows average customer sentiments. Panel B shows average agent sentiments. Panel C plots the event study of AI model deployment on customer sentiment and Panel D plots the corresponding estimate for agent sentiment. Sentiment is measured using SiEBERT, a fine-tuned checkpoint of a RoBERTA, an English language transformer model. All data come from the firm's internal software systems.

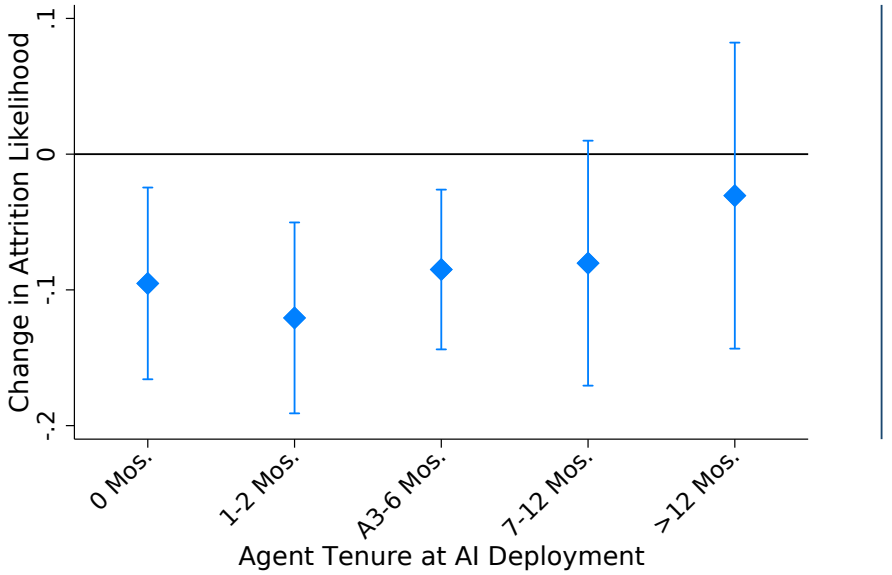
FIGURE 4.16: IMPACT OF AI ON CHAT ESCALATION



NOTES: This figure reports the coefficients and 95% confidence intervals for the event study of AI model deployment on requests for manager assistance. Standard errors are clustered at the agent level.

FIGURE 4.17: IMPACT OF AI MODEL DEPLOYMENT ON WORKER ATTRITION

A. BY TENURE AT AI MODEL DEPLOYMENT



B. BY PRODUCTIVITY AT AI MODEL DEPLOYMENT

NOTES: This figure presents the results of the impact of AI model deployment on workers' likelihood of attrition. Panel A graphs the effects of AI assistance on attrition by agent tenure at AI model deployment. Panel B plots the same impact by agent skill index at AI model deployment. All specifications include chat year and month fixed effects, as well as agent location, company and agent tenure. All robust standard errors are clustered at the agent level. All data come from the firm's internal software systems.

TABLE 4.1: APPLICANT SUMMARY STATISTICS

Variable	All	Never Treated	Treated, Pre	Treated, Post
Chats	3,007,501	945,954	882,105	1,180,446
Agents	5,179	3,523	1,341	1,636
Number of Teams	133	111	80	81
Share US Agents	.11	.15	.081	.072
Distinct Locations	25	25	18	17
Average Chats per Month	127	83	147	188
Average Handle Time (Min)	41	43	43	35
St. Average Handle Time (Min)	23	24	24	22
Resolution Rate	.82	.78	.82	.84
Resolutions Per Hour	2.1	1.7	2	2.5
Customer Satisfaction (NPS)	79	78	80	80

NOTES: This table shows conversations, agent characteristics and issue resolution rates, customer satisfaction and average call duration. The sample in Column 1 consists of all agents in our sample. Column 2 includes control agents who were never given access to the AI model. Column 3 and 4 present pre-and-post AI model deployment summary statistics for treated agents who were given access to the AI model. All data come from the firm's internal software systems.

TABLE 4.2: MAIN EFFECTS: PRODUCTIVITY (RESOLUTIONS PER HOUR)

VARIABLES	(1) Res./Hr	(2) Res./Hr	(3) Res./Hr	(4) Log(Res./Hr)	(5) Log(Res./Hr)	(6) Log(Res./Hr)
Post AI X Ever Treated	0.468***	0.371***	0.301***	0.221***	0.180***	0.138***
	(0.0542)	(0.0520)	(0.0498)	(0.0211)	(0.0188)	(0.0199)
Ever Treated	0.109*			0.0572*		
	(0.0582)			(0.0316)		
Observations	13,225	12,328	12,328	12,776	11,904	11,904
R-squared	0.250	0.563	0.575	0.260	0.571	0.592
Year Month FE	YES	YES	YES	YES	YES	YES
Location FE	YES	YES	YES	YES	YES	YES
Agent FE	-	YES	YES	-	YES	YES
Agent Tenure FE	-	-	YES	-	-	YES
DV Mean	2.121	2.174	2.174			

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on our main measure of productivity, resolutions per hour, the number of technical support problems resolved by an agent per hour (res/hour). Columns 1 and 4 include agent geographic location and year-by-month fixed effects. Columns 2 and 5 include agent-level fixed effects, and columns 3 and 6, our preferred specification, also control for agent tenure. All standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Table 4.3: Main Effects: Additional Outcomes

VARIABLES	(1) AHT	(2) Calls/Hr	(3) Res. Rate	(4) NPS	(5) Log(AHT)	(6) Log(Calls/Hr)	(7) Log(Res. Rate)	(8) Log(NPS)
Post AI X Ever Treated	-3.750***	0.366***	0.0128*	-0.128	-0.0851***	0.149***	0.00973*	-0.000406
	(0.476)	(0.0363)	(0.00717)	(0.660)	(0.0110)	(0.0142)	(0.00529)	(0.00915)
Observations	21,885	21,885	12,328	12,578	21,885	21,885	11,904	12,188
R-squared	0.590	0.564	0.369	0.525	0.622	0.610	0.394	0.565
Year Month FE	YES	YES	YES	YES	YES	YES	YES	YES
Location FE	YES	YES	YES	YES	YES	YES	YES	YES
Agent FE	YES	YES	YES	YES	YES	YES	YES	YES
Agent Tenure FE	YES	YES	YES	YES	YES	YES	YES	YES
DV Mean	40.65	2.557	0.821	79.58				

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on measures of productivity and agent performance. Post AI X Treated measures the impact of AI model deployment after deployment on treated agents for average handle time or average call duration, chats per hour, the number of chats an agent handles per hour, resolution rate, the share of technical support problems they can resolve and net promoter score (NPS), an estimate of customer satisfaction, and each metrics corresponding natural log equivalents. All specifications include agent fixed effects and chat year and month fixed effects, as well as controls for agent location and agent tenure. All standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Table 4.4: Agent and Customer Sentiment

VARIABLES	(1) Mean(Customer Sentiment)	(2) Mean(Agent Sentiment)
Post AI X Ever Treated	0.177***	0.0198***
	(0.0133)	(0.00315)
Observations	21,218	21,218
R-squared	0.485	0.596
Year Month FE	YES	YES
Location FE	YES	YES
Agent FE	YES	YES
Agent Tenure FE	YES	YES
DV Mean	0.141	0.896

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on measures of conversation sentiment. All specifications include agent fixed effects and chat year and month fixed effects, as well as agent location and agent tenure, which account for differing likelihood of attrition by agent tenure. All standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Appendix A

The Market Effects of Algorithms Appendix Materials

FIGURE A.1: CAPITALIZATION RATE EXAMPLE

DELPHI PROPERTY GROUP

PRO FORMA

Income:		
Residential		
Gross Revenue	$	778,200.00
Vacancy; 5%	$	(38,910.00)
Effective Gross Residential Income:	**$**	**739,290.00**
Commercial		
Gross Revenue	$	150,000.00
Vacancy; 5%	$	(7,500.00)
Effective Gross Commercial Income:	**$**	**142,500.00**
Total Gross Revenue	**$**	**881,790.00**

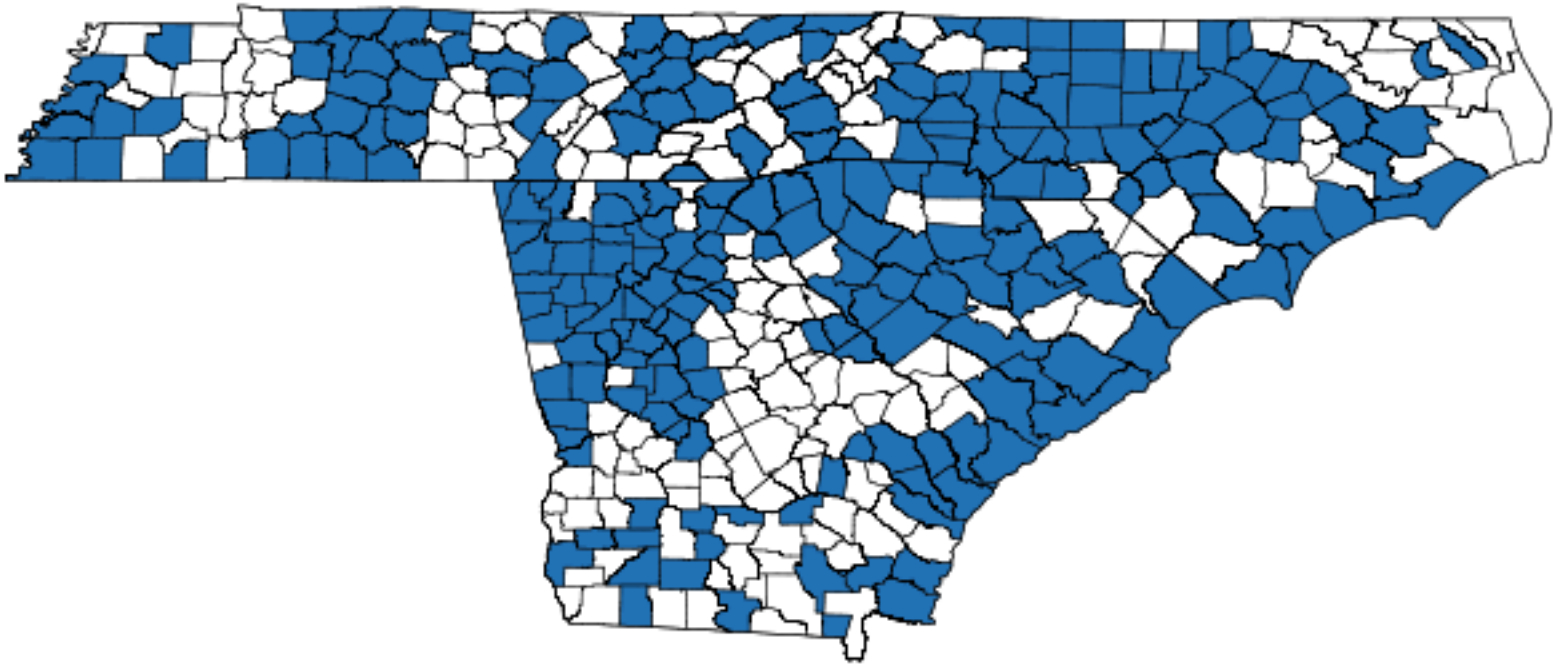
Expenses:			
Taxes		$	74,176.13
Management Fee	*5.0%*	$	44,089.50
CAM - *Estimated*		$	45,000.00
Miscallaneous - *Estimated*		$	30,000.00
Insurance		$	11,511.00
Electric (Common)		$	10,000.00
Water		$	5,000.00
Trash		$	141.60
Advanced Disposal		$	3,468.36
Total Expenses		**$**	**223,386.59**
Net Operating Income		**$**	**658,403.41**

Pricing	
Sale Price	$11,000,000.00
Number of Units	47 Apartments & 2 Commercial
Price / Unit	$224,489.80
Gross Building Area	54,000 SF
Price PSF	$203.70

Investment Summary	
Cap Rate	6.0%
NOI	$658,403.41

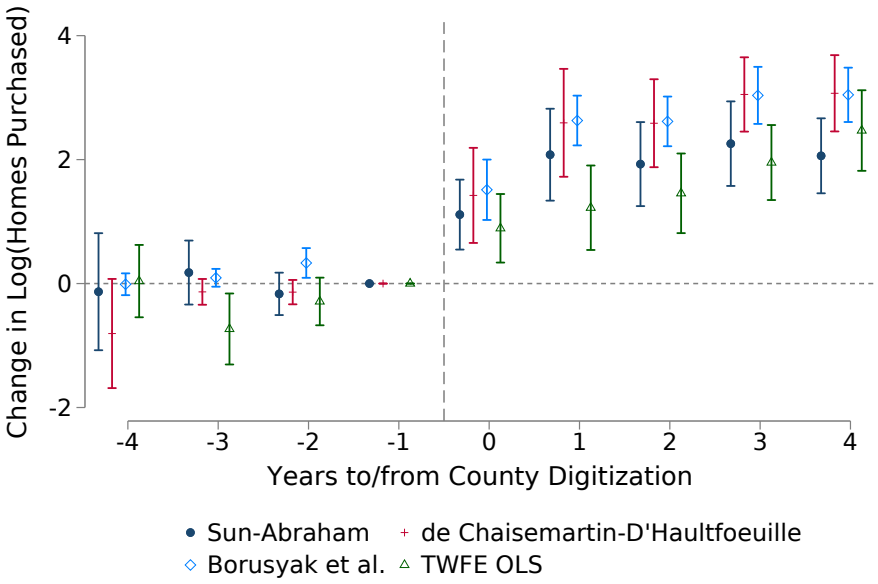
NOTES: This shows a sample of the marketing material for 1 West Main Street Norristown, PA, a mixed use multifamily apartment building. This page includes the building capitalization rate.

FIGURE A.2: INVESTOR ACTIVITY



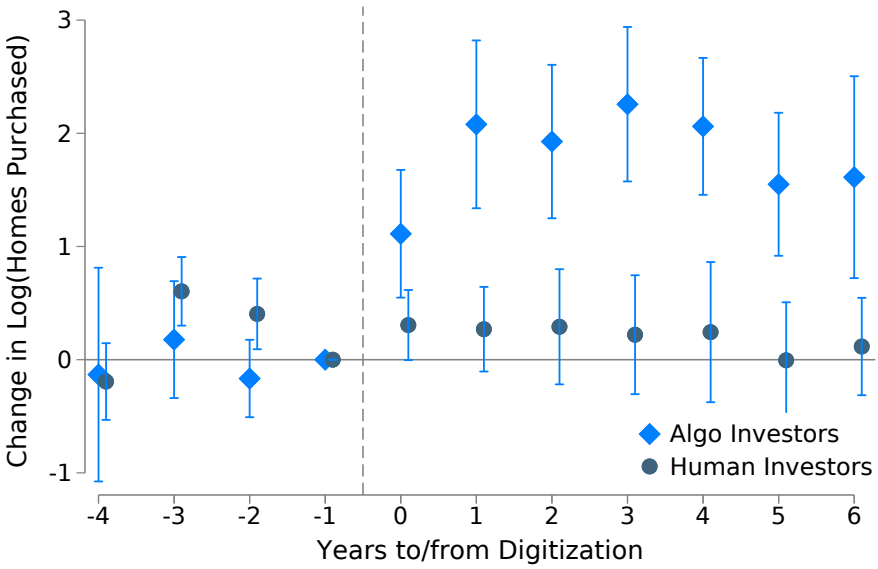
NOTES: This graph shows the counties in Georgia, North Carolina, South Carolina and Tennessee where human and algorithm investors are active in blue. Counties in White have only human investors.

FIGURE A.3: ALTERNATIVE EVENT STUDIES, LOG(HOUSES PURCHASED) BY ALGORITHMIC INVESTORS



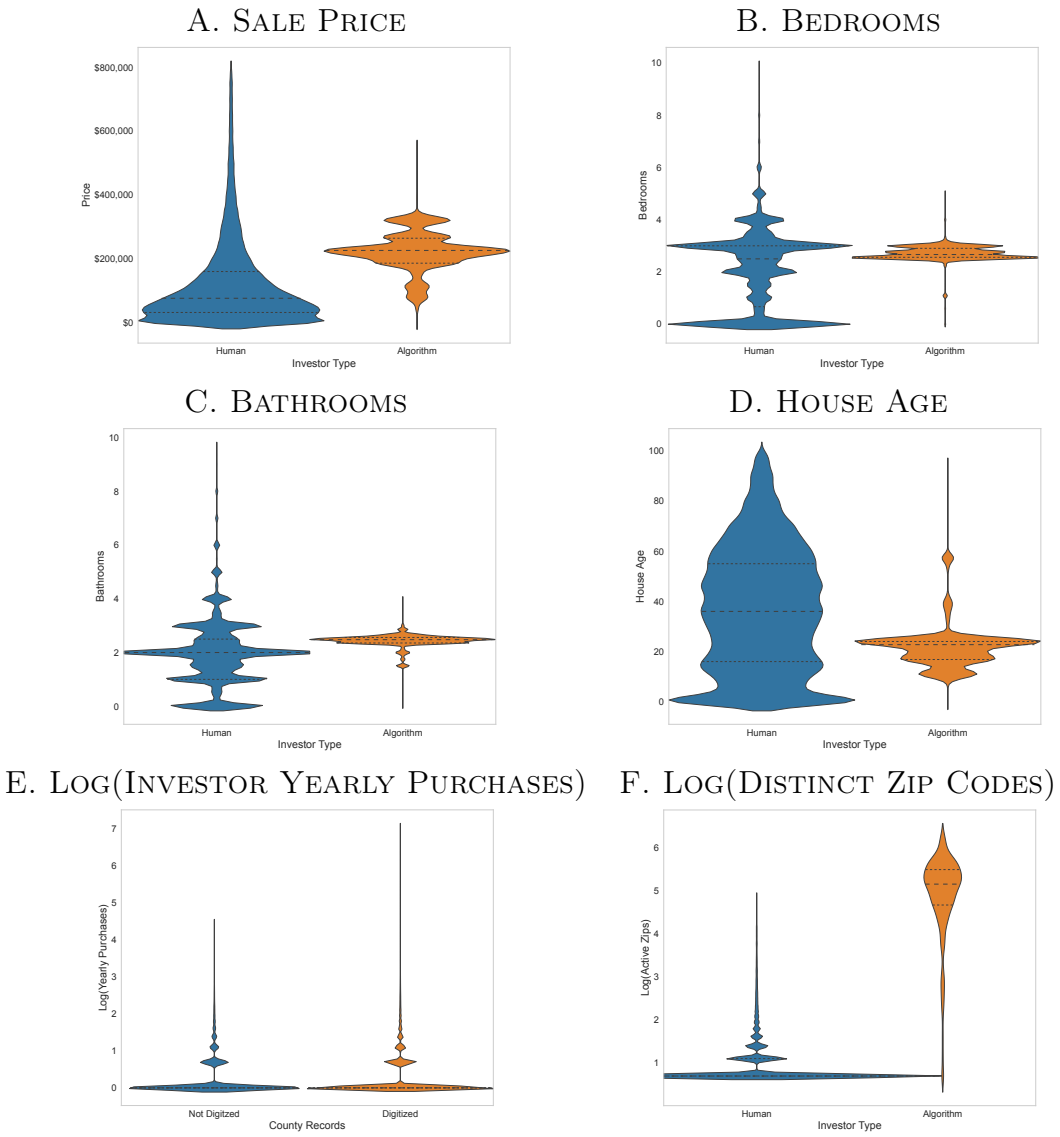
NOTES: These figures plot the coefficients and 95 percent confidence intervals using a variety of robust dynamic difference-in-differences estimators introduced in Borusyak, Jaravel and Spiess (2022), de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021) and a standard two-way fixed effects regression model. All specifications include state and year fixed effects, standard errors are clustered at the county level and are weighted by the number of transactions. All data comes from ATTOM Data, Zillow, and county digitization records.

FIGURE A.4: Log(House Purchases) by Investor Type



NOTES: This figure plots coefficients and 95 percent confidence interval from Sun and Abraham (2021) interaction-weighted event study regressions of county digitization on the natural log of the quantity of homes. I plot these results separately for the number of houses purchased by human or algorithmic investors in each county and year, weighted by the number of transactions. The regression includes state and year fixed effects, and standard errors are clustered at the county level. All data comes from ATTOM Data, Zillow and county digitization records.

Figure A.5: House Characteristics, by Investor



Notes: This figures plots characteristics of houses purchased by human and algorithmic investors. Panel A plots the purchase prices of houses, Panel B plots the number of house bedrooms. Panel C shows the number of bathrooms and panel D shows the age of the house. Panel E plots the natural log of average houses purchased by investors each before and after digitization. Panel F plots the natural log of zip codes investors are active each year. All data come from ATTOM Data and Zillow.

FIGURE A.6: HOUSE EXTERIOR IMAGES

NOTES: This shows an example of the exterior images of the house used in the deep learning model.

TABLE A.1: BALANCE TABLE: COUNTIES, BY YEAR OF DIGITIZATION

Variable	(1) Early Digitizers	(2) Late Digitizers	(3) Difference
Population	84,157.59	49,187.26	-34,970.33***
	(144,805.80)	(51,842.46)	(0.00)
Unemployment Rate	4.69	4.51	-0.19
	(1.81)	(1.69)	(0.36)
Share in Labor Force	56.67	54.77	-1.89**
	(7.01)	(6.24)	(0.01)
Share Units Occupied	82.58	81.33	-1.25
	(8.68)	(8.82)	(0.23)
Share Vacant	2.17	1.90	-0.27
	(1.44)	(2.17)	(0.27)
Median Rent	710.86	679.33	-31.53*
	(164.93)	(161.44)	(0.10)
Share Families in Poverty	14.66	14.62	-0.04
	(5.25)	(4.99)	(0.95)
Mean Family Size	3.14	3.07	-0.07***
	(0.29)	(0.20)	(0.01)
Median Income	44,399.43	42,521.30	-1,878.12
	(11,331.64)	(12,210.29)	(0.19)
Share Black	22.99	19.92	-3.07
	(18.08)	(19.44)	(0.17)
Share Hispanic	5.81	4.63	-1.18***
	(4.61)	(3.64)	(0.01)
Share White	67.19	71.74	4.56*
	(19.80)	(20.63)	(0.06)
Share Asian	1.25	0.85	-0.40***
	(1.31)	(0.95)	(0.00)
Observations	303	97	400

NOTES: This table shows the covariate balance table for counties digitized before and after the median. All variables are calculated at the county level. All data come from ATTOM Data, ZTRAX and the US Census.

TABLE A.2: BALANCE TABLE: HOUSES, BY YEAR OF DIGITIZATION

Variable	(1) Early Digitizers	(2) Late Digitizers	(3) Difference
Years since Sale	10.55	9.79	-0.76***
	(9.49)	(8.74)	(0.00)
Sale Price	210,262.55	202,658.52	-7,604.03***
	(959,260.94)	(804,228.75)	(0.00)
Bedrooms	2.19	2.07	-0.12***
	(1.68)	(3.29)	(0.00)
Bathrooms	2.03	2.14	0.11***
	(2.69)	(2.24)	(0.00)
Partial Baths	0.29	0.27	-0.02***
	(0.52)	(0.47)	(0.00)
Stories	1.17	1.26	0.09***
	(0.88)	(0.69)	(0.00)
Buildings	0.05	0.07	0.01***
	(0.42)	(0.53)	(0.00)
Garage	0.55	0.56	0.02***
	(0.50)	(0.50)	(0.00)
Fireplace	0.60	0.58	-0.02***
	(0.49)	(0.49)	(0.00)
Basement	0.18	0.17	-0.01***
	(0.38)	(0.37)	(0.00)
Parking Spaces	0.97	0.69	-0.28***
	(17.96)	(1.77)	(0.00)
House Age	33.12	30.18	-2.94***
	(24.85)	(26.03)	(0.00)
Age Since Remodel	27.87	23.68	-4.19***
	(21.54)	(21.13)	(0.00)
Minority Homeowner	0.04	0.04	-0.00***
	(0.20)	(0.19)	(0.00)
Homeowner Asian	0.02	0.03	0.01***
	(0.15)	(0.17)	(0.00)
Homeowner White	0.88	0.87	-0.01***
	(0.33)	(0.34)	(0.00)
Observations	1,096,423	3,684,075	4,780,498

NOTES: This table shows the covariate balance table for houses that digitized before and after the median ("Early Digitizers") or later ("Late Digitizers"). The unit of observation is at the house level. All data come from ATTOM Data, ZTRAX and the US Census.

TABLE A.3: BALANCE TABLE: NEIGHBORHOOD CHARACTERISTICS, BY YEAR OF DIGITIZATION

Variable	(1) Early Digitizers	(2) Late Digitizers	(3) Difference
Population	2,065.07	2,215.21	150.13***
	(1,284.76)	(1,448.16)	(0.00)
Housing Units	928.23	992.07	63.84***
	(540.46)	(600.80)	(0.00)
Share White	68.00	67.66	-0.34***
	(26.67)	(28.04)	(0.00)
Share Black	20.04	21.25	1.21***
	(21.69)	(24.18)	(0.00)
Share Asian	2.50	2.85	0.35***
	(3.12)	(4.33)	(0.00)
Share Under 18	23.66	24.01	0.35***
	(6.34)	(6.22)	(0.00)
Median Earnings	53,533.62	54,046.14	512.52***
	(13,100.19)	(13,607.33)	(0.00)
Rent	864.28	878.85	14.57***
	(200.30)	(199.76)	(0.00)
Age	38.23	38.09	-0.13***
	(4.39)	(4.43)	(0.00)
Mortgage Costs	1,310.06	1,332.71	22.65***
	(236.65)	(256.49)	(0.00)
Median List Price	216,445.45	205,233.47	-11,211.99***
	(75,300.37)	(70,603.58)	(0.00)
Days on the Market	109.73	107.64	-2.08***
	(30.66)	(27.31)	(0.00)
Observations	1,096,423	3,684,075	4,780,498

NOTES: This table shows the covariate balance table for houses that digitized before and after the median ("Early Digitizers") or later ("Late Digitizers"). When possible, all statistics are at the census block group level. Information from Zillow is at the zip code level and the unit of observation is at the house level. All data come from ATTOM Data, ZTRAX and the US Census.

Table A.4: County Digitization and Algorithmic Investors Buying

VARIABLES	(1) Ln(Q_Algo)	(2) Ln(Q_Algo)	(3) Ln(Q_Algo)
County Digitization	1.130**	0.780**	0.749**
	(0.380)	(0.221)	(0.229)
Observations	3,962	3,962	3,962
R-squared	0.798	0.812	0.816
Year FE	Yes	Yes	Yes
Location FE	Yes	Yes	Yes
SocioEconomics	-	Yes	Yes
Housing Stock	-	-	Yes
DV Mean	2.597	2.597	2.597

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the results of county-level difference-in-difference regressions estimating the effect of county record digitization on the natural log of houses purchased by algorithmic investors. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the county level. Column 2 includes county population, demographics, poverty, unemployment rate and educational characteristics. Column 3 add housing stock characteristics such as the number of housing units and rent burden. All data comes from ATTOM Data, ZTRAX the US Census and county governments.

TABLE A.5: ALGORITHMIC INVESTOR PURCHASE, BY HOMEOWNER RACE, INVESTOR SAMPLE

	(1) Algorithm Purchase	(2) Algorithm Purchase	(3) Algorithm Purchase
Seller Minority	-0.0133***	-0.0137***	-0.0194***
	(0.0036)	(0.0036)	(0.0045)
Digitization x Seller White	0.0079*	0.0075**	0.0042*
	(0.0045)	(0.0035)	(0.0022)
Digitization x Seller Minority	0.0415***	0.0396***	0.0389***
	(0.0043)	(0.0042)	(0.0050)
Geography FE	Tract	Block Group	Block
Year FE	Yes	Yes	Yes
Sample	Investors	Investors	Investors
DV Mean	.0018	.0018	.0018
Observations	898975	898061	802192

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

NOTES: This table shows the results of cross-sectional difference-in-difference regressions estimating the impact of house record digitization on the purchase by an algorithmic investor, by homeowner race. The sample includes only investor purchases, so the coefficients are interpreted as the likelihood of being purchased by an algorithmic investor compared to human investors. All specifications include house characteristics, year, and geography fixed effects, and standard errors are clustered at the geographic level. All data comes from ATTOM Data, ZTRAX and county governments.

TABLE A.6: IV ANALYSIS: ALGORITHMIC INVESTORS AND RACE PENALTY

	(1) First Stage	(2) 2SLS	(3) First Stage	(4) 2SLS	(5) First Stage	(6) 2SLS
Digitization	0.043***		0.046***		0.067***	
	(0.004)		(0.004)		(0.006)	
Algo Buyer		0.289		0.279		0.291
		(0.434)		(0.320)		(0.209)
AlgoxSellerBlack/Hisp		0.526***		0.529***		0.527***
		(0.127)		(0.116)		(0.107)
Geo Level	Tract+Year	Tract+Year	BG+Year	BG+Year	Block+Year	Block+Year
DV Mean	.002	164167	.002	164167	.002	164167
Adj R-squared	.317	.345	.317	.345	.344	.345
Observations	222666	222772	221537	222686	151452	222686

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table shows the results of cross-sectional 2SLS regressions that estimate the algorithmic investor purchase on the race penalty, instrumenting for the algorithmic purchase with house-level digitization. All specifications include house characteristics, year and geography fixed effects, and standard errors are clustered at the geographic level and use log sale price as the outcome. All data comes from ATTOM Data, ZTRAX and county governments.

Table A.7: Assessment Margin

VARIABLES	(1) Log(Assess Margin)	(2) Log(Assess Margin)	(3) Log(Assess Margin)	(4) Log(Assess Margin)	(5) Log(Assess Margin)	(6) Log(Assess Margin)	(7) Log(Assess Margin)	(8) Log(Assess Margin)
Seller Minority = 1	0.004	0.004	0.002	0.003	-0.013*	-0.005	0.022**	-0.052***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.007)	(0.008)	(0.010)	(0.012)
Seller Minority x Minority Neighborhood = 1				0.003				0.090***
				(0.006)				(0.015)
Observations	81,387	76,312	46,294	76,312	467,588	440,142	238,501	440,142
R-squared	0.714	0.748	0.820	0.748	0.362	0.470	0.723	0.470
FE	Year x Tract	Year x Block Group	Year x Block	Year x Block Group	Year x Tract	Year x Block Group	Year x Block	Year x Block Group
Assessment Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Buyers	Algorithms	Algorithms	Algorithms	Algorithms	Humans	Humans	Humans	Humans
DV Mean	0.0911	0.0839	0.0577	0.0839	0.698	0.703	0.795	0.703

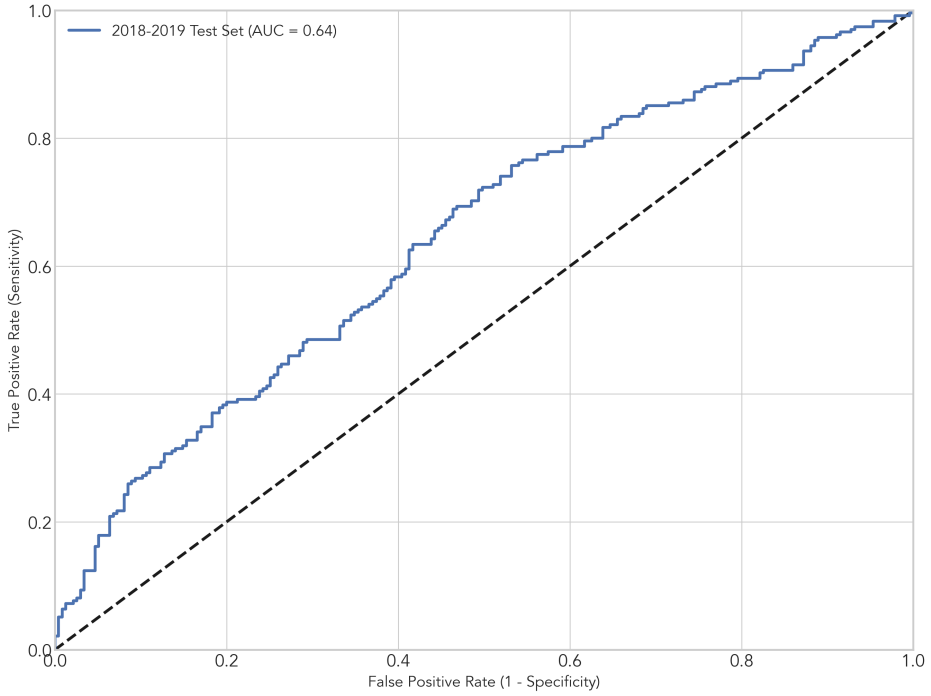
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the difference in the natural log of estimated house market value and the natural log transformation of the price paid, or *assessment margin*. *Minority neighborhood* indicates if the house is in a census block group with an above average minority resident share. All specifications include tax estimate and sale year by geography fixed effects. Standard errors are clustered at the geographic level. All data comes from ATTOM Data.

Appendix B

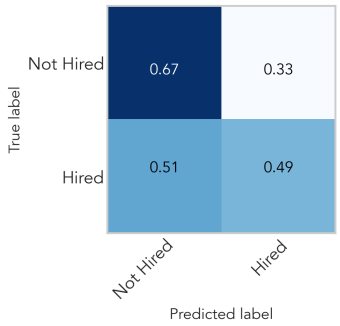
Hiring as Exploration Appendix Materials

FIGURE A.1: MODEL PERFORMANCE: PREDICTING HIRING, CONDITIONAL ON RECEIVING AN INTERVIEW



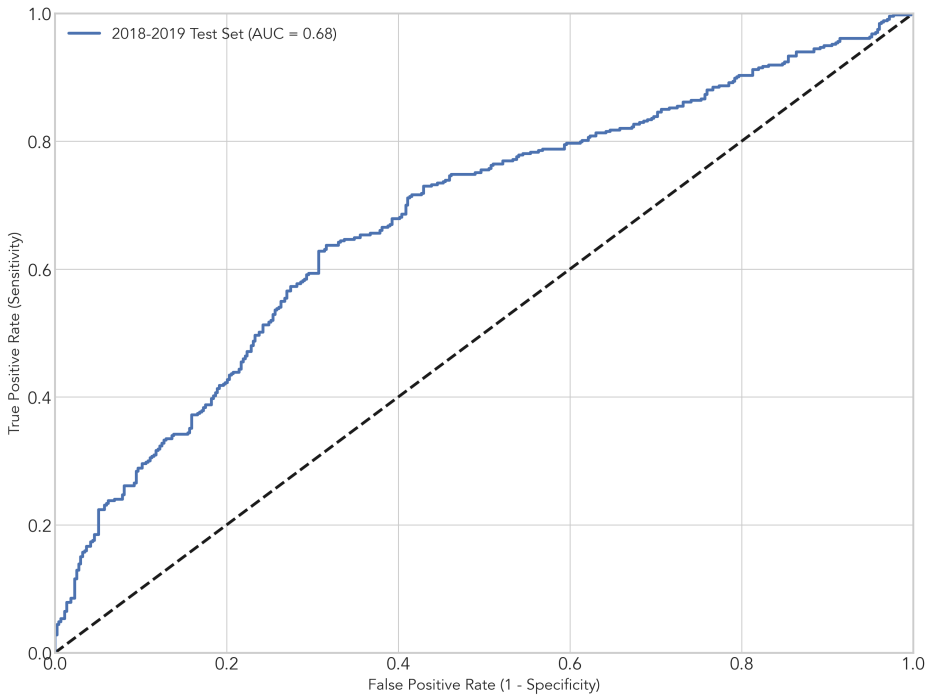
NOTES: This figure shows the Receiver-Operating Characteristic (ROC) curve for the baseline supervised learning model, which predicts hiring potential. The ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis. For reference, the 45 degree line is shown with a black dash in each plot. All data come from the firm's application and hiring records.

FIGURE A.2: CONFUSION MATRIX MODEL PERFORMANCE: PREDICTING HIRING, CONDITIONAL ON RECEIVING AN INTERVIEW



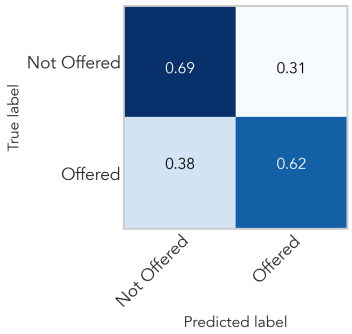
NOTES: This figure shows a confusion matrix for the baseline supervised learning model, which predicts hiring potential. The confusion plots the predicted label on the x-axis and the true label rate on the y-axis. Correctly classified applicants are in the top left cell, "true positives" and bottom right, "true negatives". Examples that are incorrectly classified are in the top left cell ("false positives") and the bottom right ("false negatives"). All data come from the firm's application and hiring records.

FIGURE A.3: MODEL PERFORMANCE: PREDICTING OFFER, CONDITIONAL ON RECEIVING AN INTERVIEW



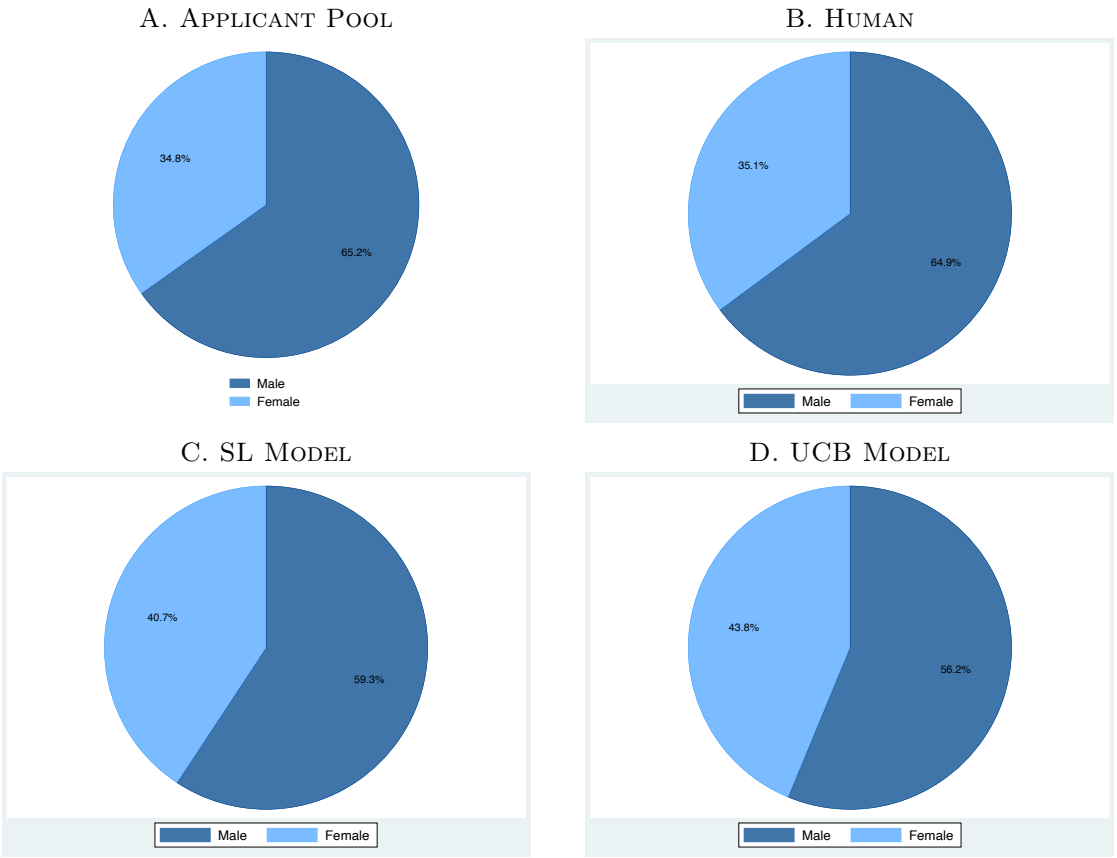
NOTES: This figure shows the Receiver-Operating Characteristic (ROC) curve for the baseline supervised learning model, which predicts offer potential. The ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis. For reference, the 45 degree line is shown with a black dash in each plot. All data come from the firm's application and hiring records.

FIGURE A.4: CONFUSION MATRIX MODEL PERFORMANCE: PREDICTING OFFER, CONDITIONAL ON RECEIVING AN INTERVIEW



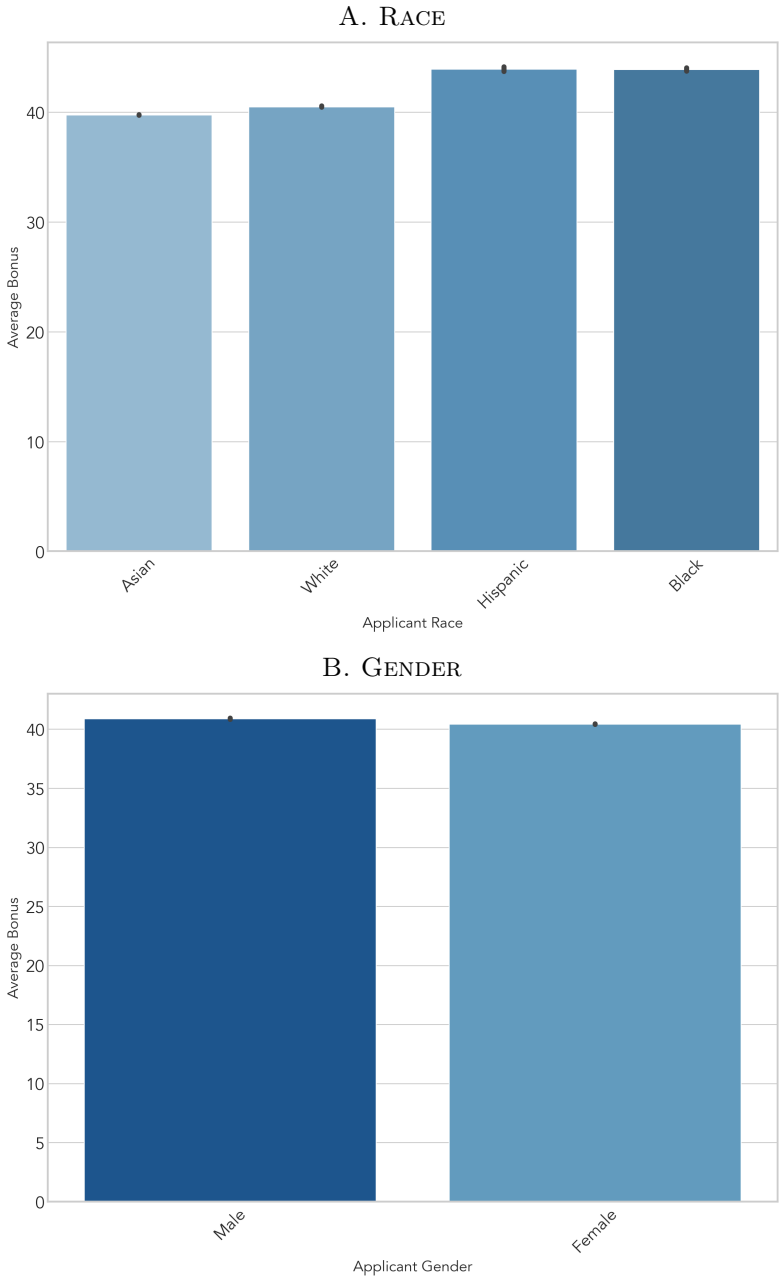
NOTES: This figure shows a confusion matrix for the baseline supervised learning model, which predicts offer potential. The confusion plots the predicted label on the x-axis and the true label rate on the y-axis. Correctly classified applicants are in the top left cell, "true positives" and bottom right, "true negatives". Examples that are incorrectly classified are in the top left cell ("false positives") and the bottom right ("false negatives"). All data come from the firm's application and hiring records.

Figure A.5: Gender Composition



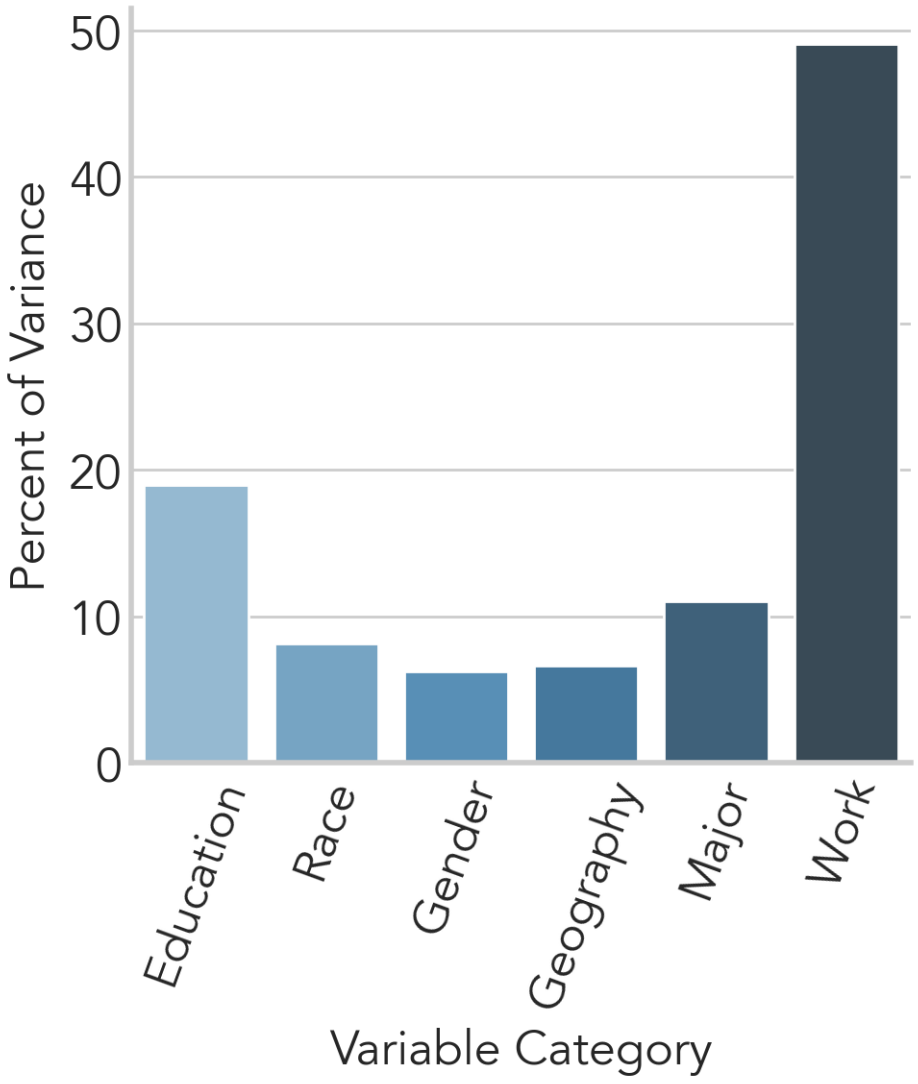
Notes: Panel A shows the gender composition of applicants in our data. Panel B shows the composition of applicants actually selected for an interview by the firm. Panel B shows the racial composition of applicants who would be selected if chosen by the supervised learning algorithm described in Equation (3.5). Finally, Panel D shows the composition of applicants who would be selected for an interview by the UCB algorithm described in Equation (3.6). All data come from the firm's application and hiring records.

Figure A.6: UCB Bonuses



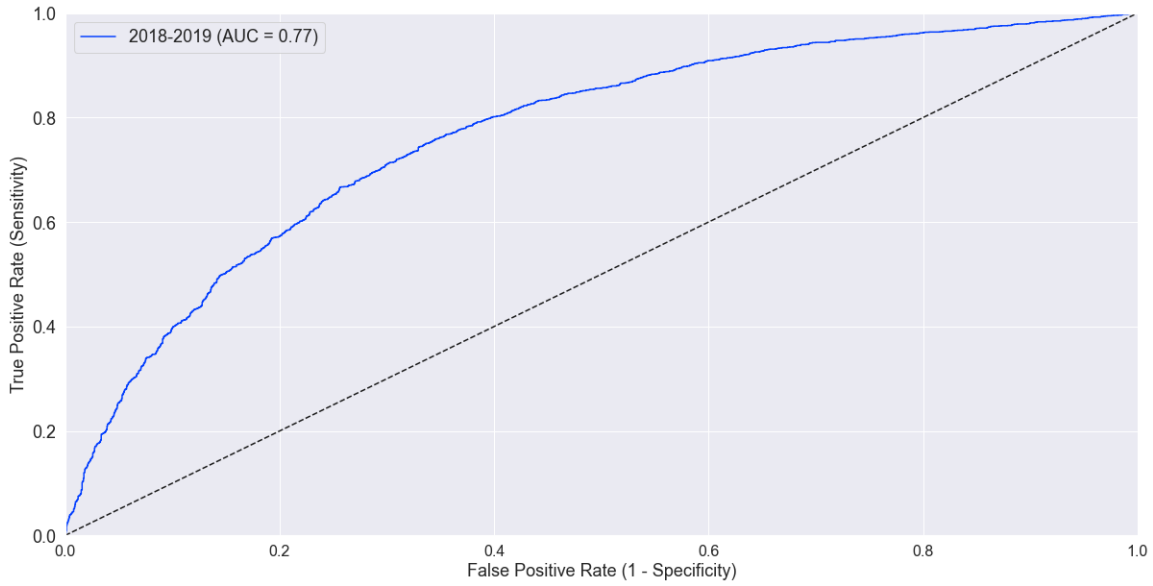
Notes: This figure shows UCB exploration bonuses averaged over the testing period. Panel A focuses on race while Panel B focuses on gender.

FIGURE A.7: DRIVERS OF VARIATION IN EXPLORATION BONUSES



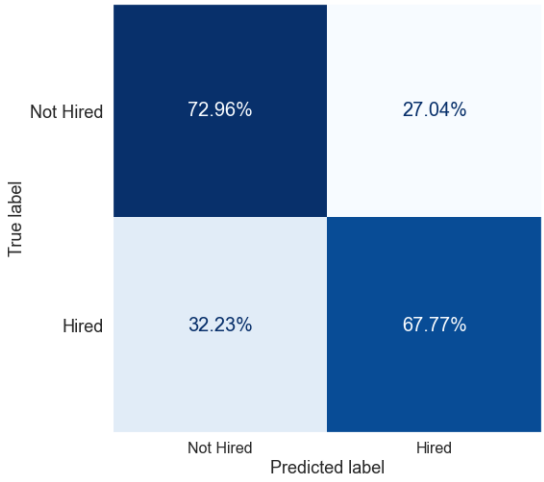
NOTES: This figure shows the percent of applicant covariate-driven variation in exploration bonuses associated with various categories of applicant features. Education refers to information such as college degree and ranking of college attended. Geography captures the geographic location of educational experience, such as India, China or the US. Major includes the coding of majors for each educational degree above high school. Work includes information on previous work experience, such as whether an applicant has experience in a Fortune 500 firm. The interactions category includes race and gender by degree and ranking of college or university.

FIGURE A.8: MODEL PERFORMANCE: PREDICTING INTERVIEW SELECTION



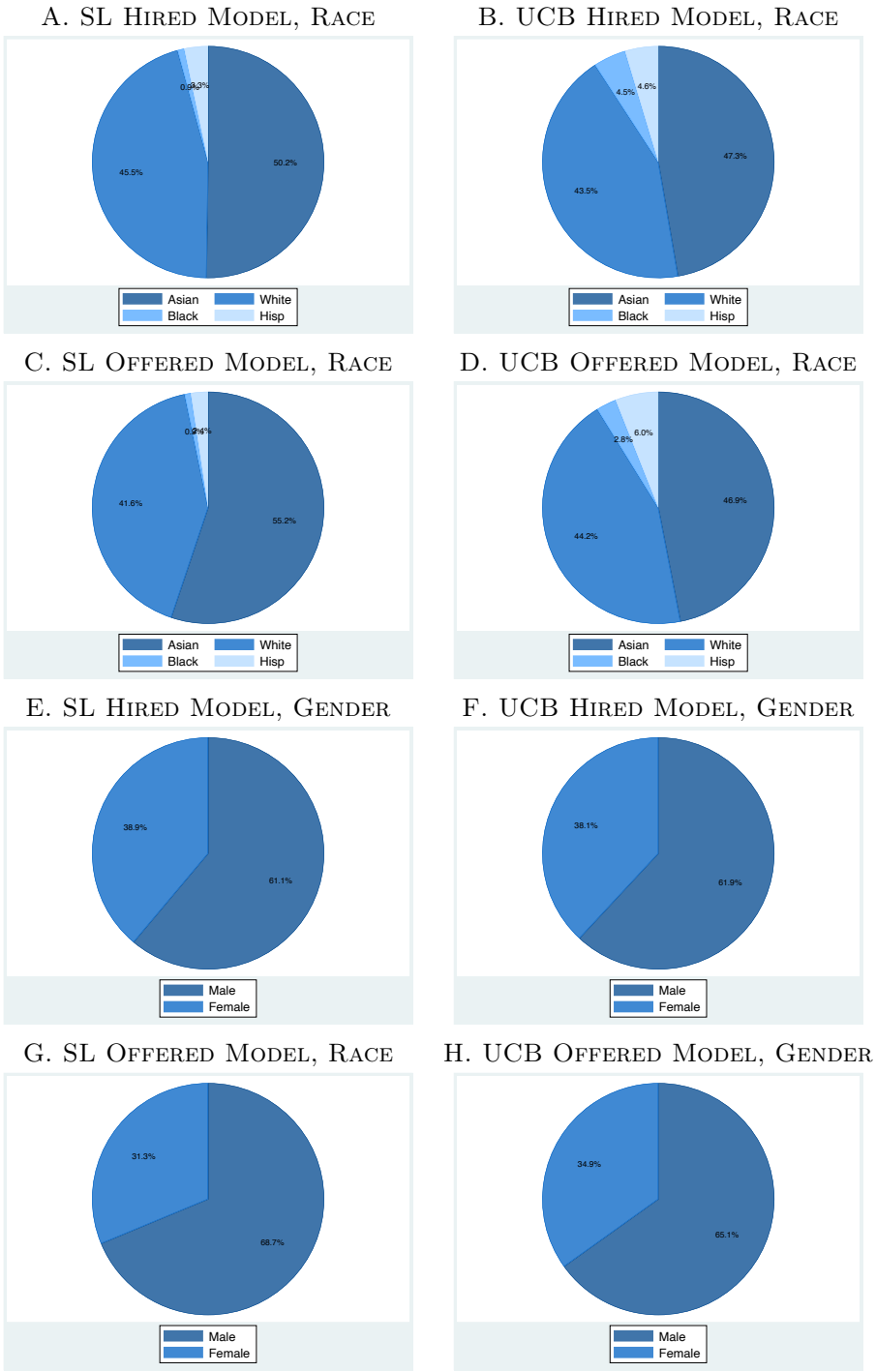
NOTES: This figure shows Receiver-Operating Characteristic (ROC) curve for the human decision making model, which is trained to predict an applicant's likelihood of being selected for an interview. The ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis. For reference, the 45 degree line is shown with a black dash in each plot. All data come from the firm's application and hiring records.

FIGURE A.9: CONFUSION MATRIX MODEL PERFORMANCE: PREDICTING INTERVIEW SELECTION



NOTES: This figure shows a confusion matrix for the human decision making model, which is trained to predict an applicant's likelihood of being selected for an interview. The confusion plots the predicted label on the x-axis and the true label rate on the y-axis. Correctly classified applicants are in the top left cell, "true positives" and bottom right, "true negatives". Examples that are incorrectly classified are in the top left cell ("false positives") and the bottom right ("false negatives"). All data come from the firm's application and hiring records.

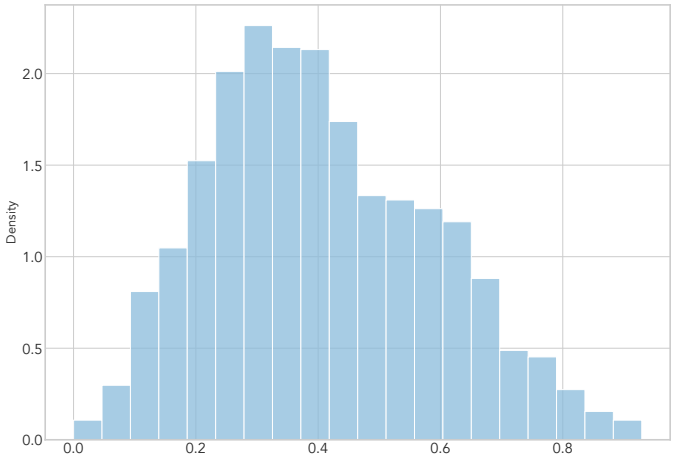
Figure A.10: Demographic Diversity: Selecting Top 50% Among Interviewed



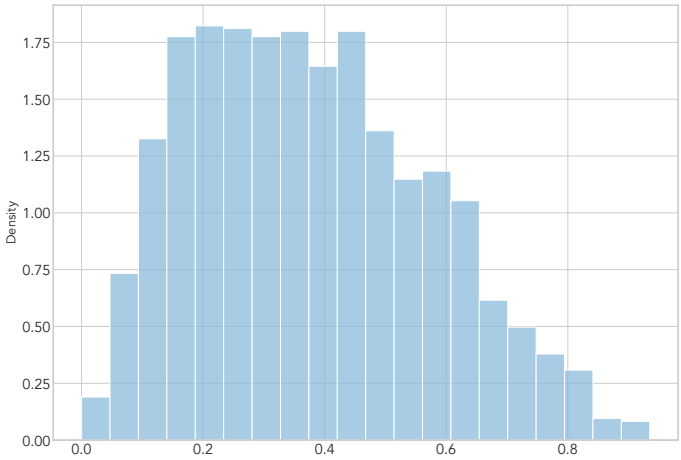
Notes: These panels consider the demographic diversity of candidates, selecting amongst the interviewed candidates. Here, we consider the scenario in which we select the top half of candidates as ranked by each ML score. Results are similar if we use other selection rules. All data come from the firm's application and hiring records.

FIGURE A.11: DISTRIBUTION OF HUMAN SELECTION PROPENSITY, AMONG ML-SELECTED APPLICANTS

A. SL SELECTED CANDIDATES



B. UCB SELECTED CANDIDATES

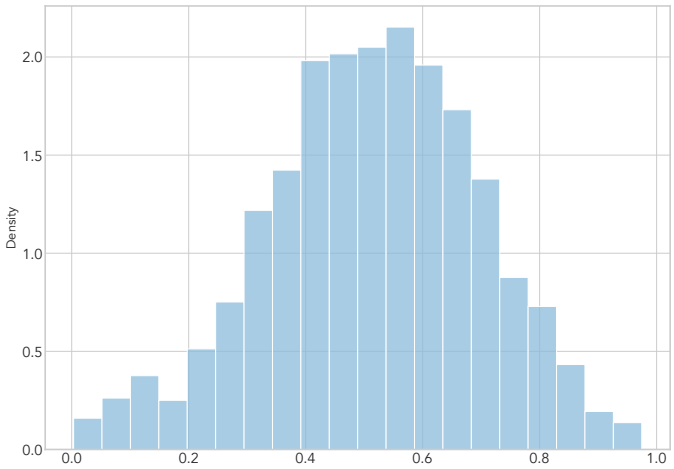


NOTES: This figure shows the distribution of propensity scores for selection into the interview set, $p(I = 1|X)$, by human recruiters under the SL and UCB models. In each panel, we plot the distribution of the propensity scores for set of applicants selected by I^{SL} and I^{UCB}.

NOTES: This figure shows the distribution of propensity scores for selection into the interview set, $p(I = 1|X)$, by human recruiters under the SL and UCB models. In each panel, we plot the distribution of the propensity scores for set of applicants selected by I^{SL} and I^{UCB}.

Figure A.12: Distribution of Human Selection Propensity, among ML-selected Applicants, Offer Model

A. SL Selected Candidates



B. UCB Selected Candidates

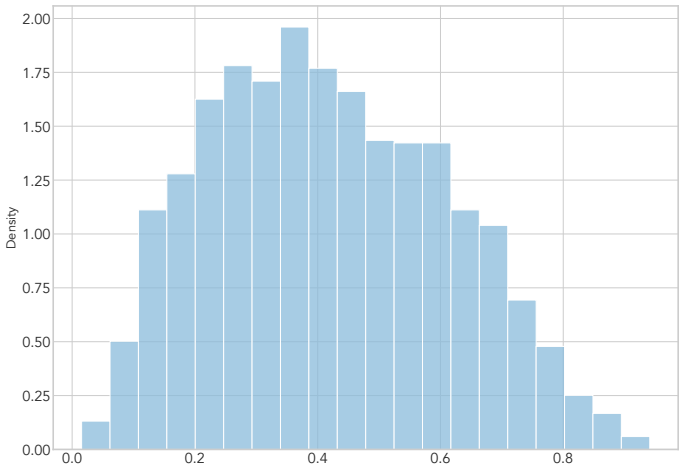
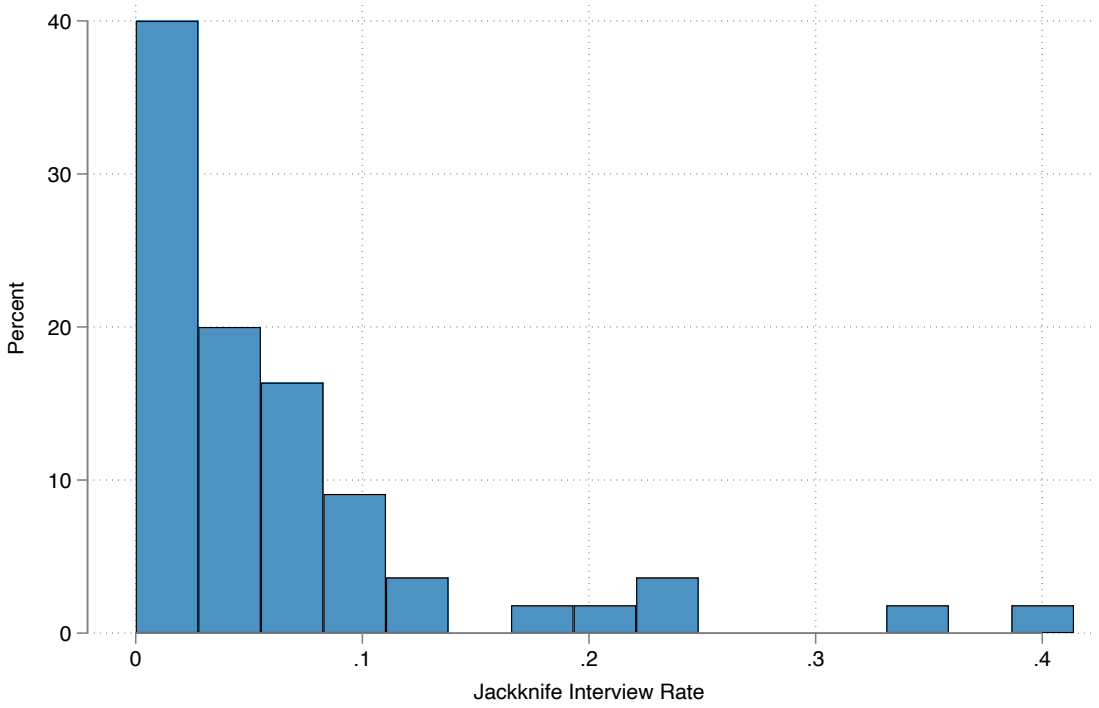


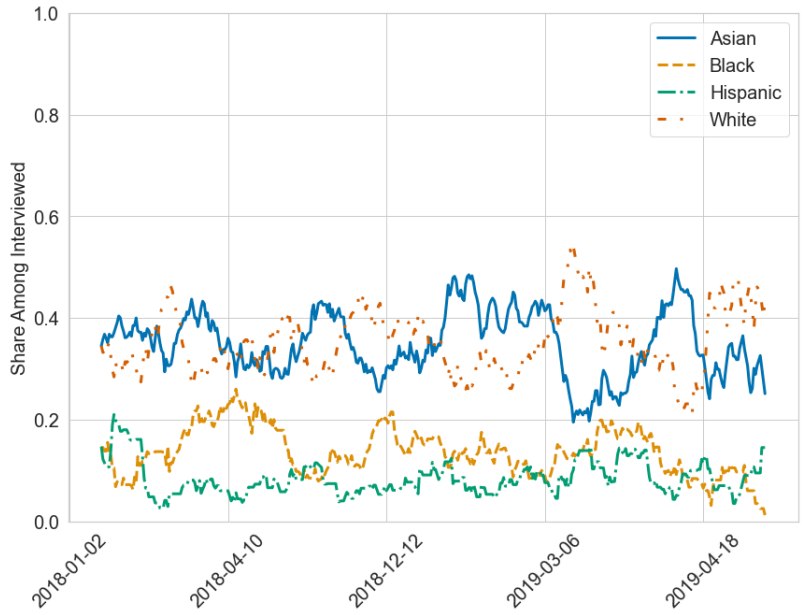
FIGURE A.13: DISTRIBUTION OF INTERVIEW RATES



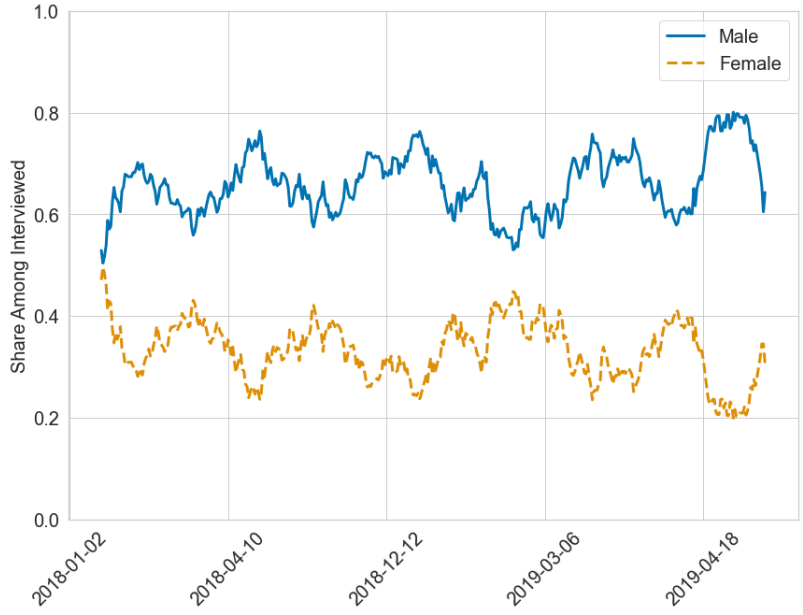
NOTES: This histogram shows the distribution of jack-knife interview rates for the 54 screeners in our data who evaluate more than 50 applicants. All data come from the firm's application and hiring records.

FIGURE A.14: UCB COMPOSITION OF SELECTED CANDIDATES, OVER TIME

A. RACE



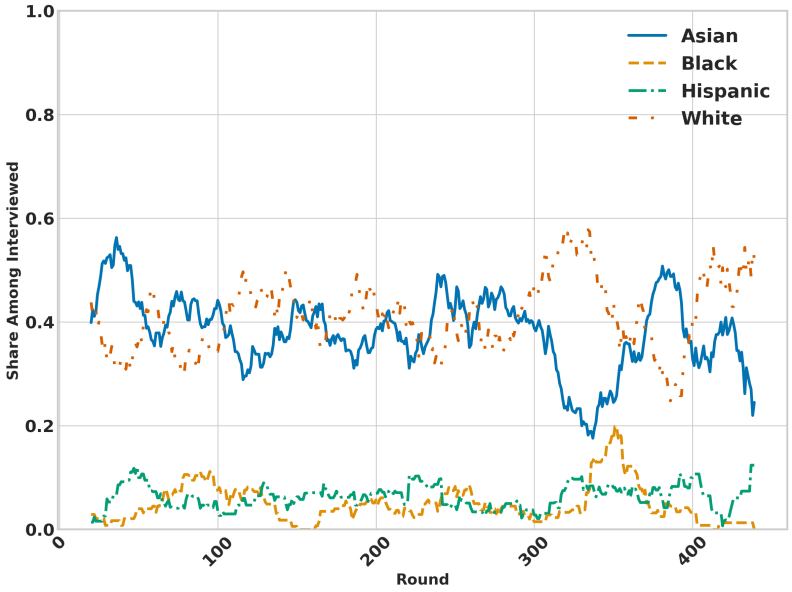
B. GENDER



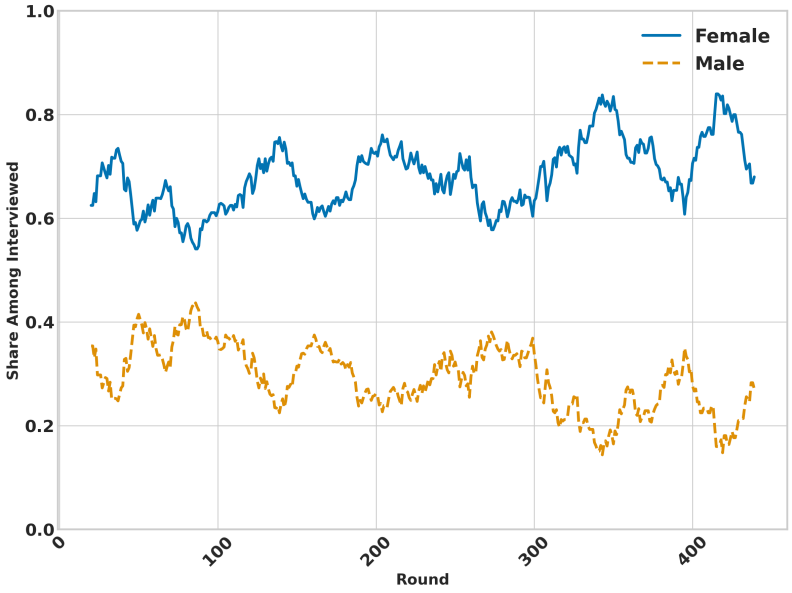
NOTES: This figure shows the composition of applicants selected to be interviewed by the UCB model at each point during the analysis period. Panel A focuses on race while Panel B focuses on gender.

FIGURE A.15: UCB COMPOSITION OF SELECTED CANDIDATES, OVER TIME, OFFER MODEL

A. RACE

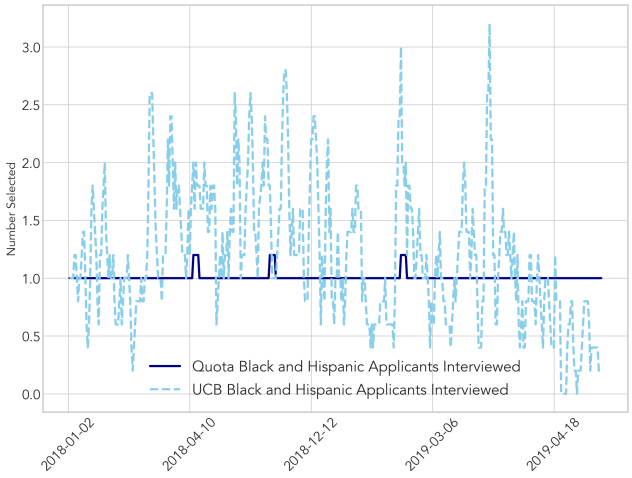


B. GENDER



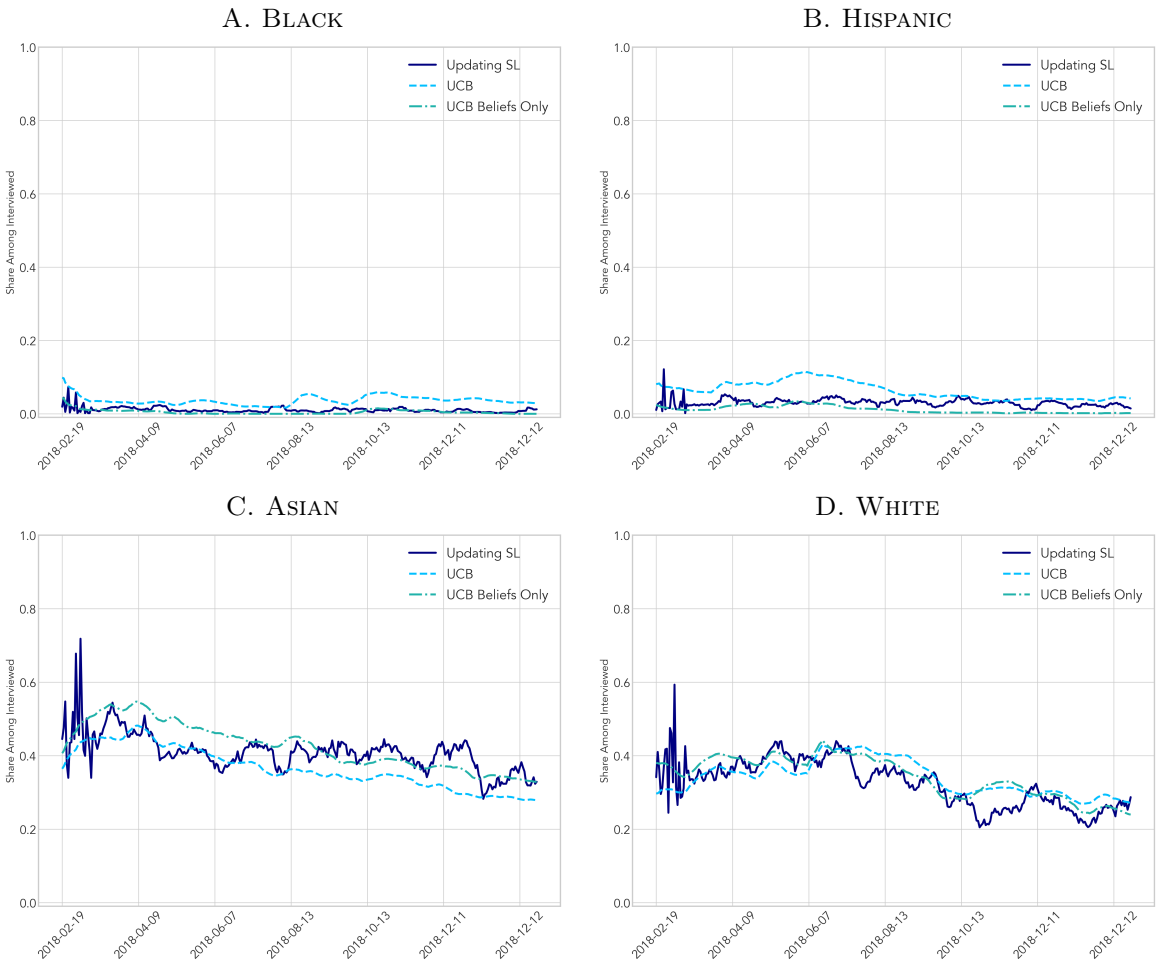
NOTES: This figure shows the composition of applicants selected to be interviewed by the UCB model at each point during the analysis period. Panel A focuses on race while Panel B focuses on gender.

FIGURE A.16: NUMBER OF BLACK OR HISPANIC CANDIDATES SELECTED, UCB VERSUS SL WITH QUOTA



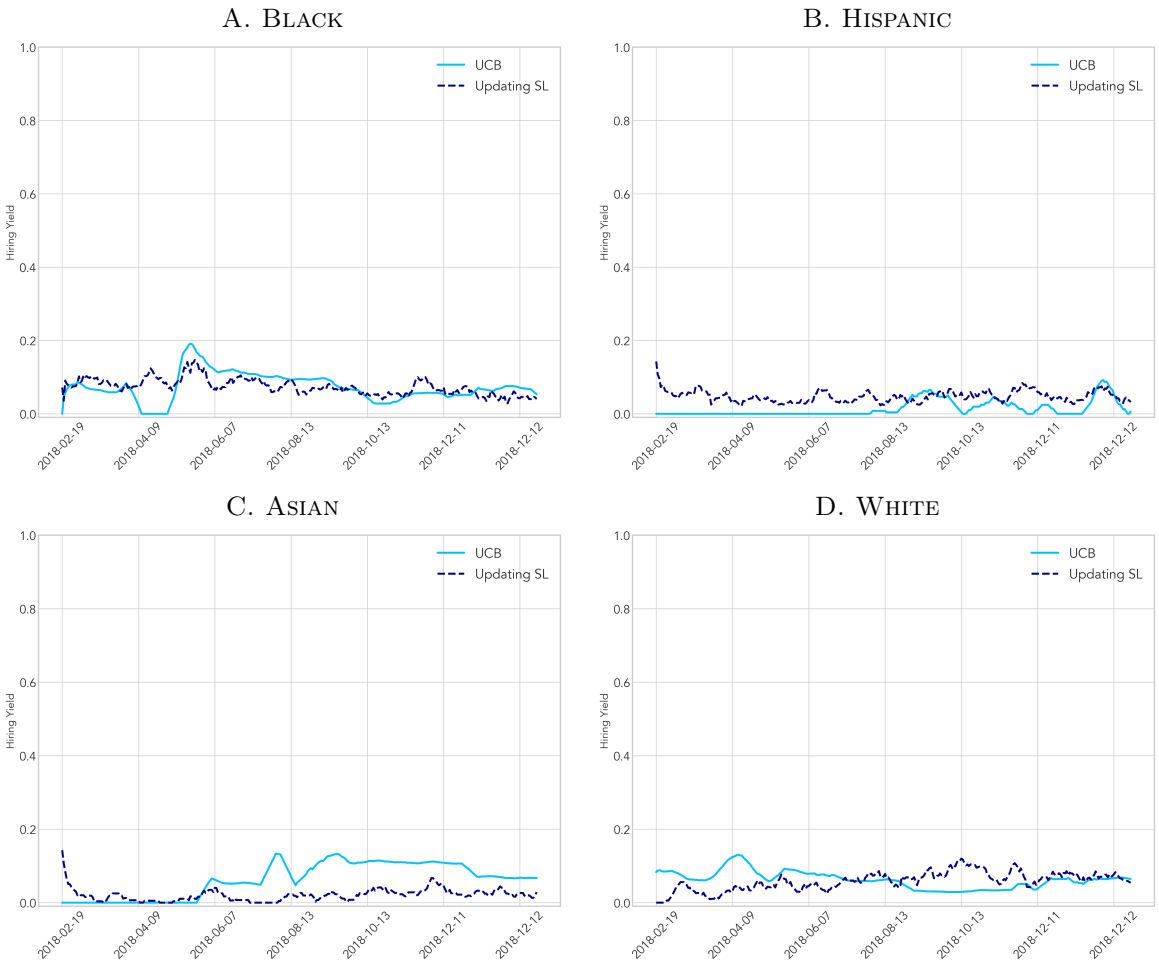
NOTES: This figure plots the number of Black or Hispanic applicants selected in each round for the baseline UCB and SL with quota models.

Figure A.17: Dynamic Updating, Decreased Quality



Notes: This figure shows the share of applicants recommended for interviews under three different algorithmic selection strategies: SL, UCB, and the beliefs component of UCB (that is, the $\hat{E}_t[H|X; D_t^{UCB}]$ term in Equation (3.6)). In each panel, the y-axis graphs the share of "evaluation cohort" (2019) applicants who would be selected under each simulation. Panel A plots the share of evaluation cohort Black applicants who would be selected under the simulation in which the hiring potential of Black candidates decreases linearly over the course of 2018, to $H = 0$. Panel B shows results from a simulation in which the hiring potential of Hispanic candidates in 2018 decreases in the same manner. Similarly, Panels C and D show results from simulations in which the hiring potential of White and Asian applicants decreases, respectively.

FIGURE A.18: DYNAMIC UPDATING, DECREASED QUALITY, ACCURACY



NOTES: This figure shows the share of applicants recommended for interviews that are also hired by humans under three different algorithmic selection strategies: SL, UCB, and the beliefs component of UCB (that is, the $\hat{E}_t[H|X; D_t^{UCB}]$ term in Equation (3.6)). In each panel, the y-axis graphs the share of "evaluation cohort" (2019) applicants who would be selected under each simulation who are also hired by humans. Panel A plots the share of evaluation cohort Black applicants who would be selected under the simulation in which the hiring potential of Black candidates decreases linearly over the course of 2018, to $H = 0$. Panel B shows results from a simulation in which the hiring potential of Hispanic candidates in 2018 decreases in the same manner. Similarly, Panels C and D show results from simulations in which the hiring potential of White and Asian applicants decreases, respectively.

TABLE A.1: APPLICANT FEATURES AND SUMMARY STATISTICS

Variable	Mean Training	Mean Test	Mean Overall
Worked at a Fortune 500 Co.	0.02	0.02	0.02
Has a Quantitative Background	0.23	0.27	0.25
Attended School in China	0.07	0.08	0.08
Attended School in Europe	0.05	0.05	0.05
Attended School in India	0.21	0.24	0.22
Attended School in Latin America	0.01	0.01	0.01
Attended School in Middle East/Africa	0.01	0.02	0.02
Attended School in Other Asian Country	0.02	0.02	0.02
Attended Elite International School	0.09	0.10	0.10
Attended US News Top 25 Ranked College	0.14	0.14	0.14
Attended US News Top 50 Ranked College	0.27	0.28	0.28
Military Experience	0.04	0.04	0.04
Number of Applications	3.5	3.8	3.5
Number of Unique Degrees	1.7	1.75	1.7
Number of Work Histories	3.8	4.0	3.9
Has Service Sector Experience	0.01	0.01	0.01
Major Description Business Management	0.17	0.15	0.17
Major Description Computer Science	0.14	0.13	0.14
Major Description Finance/Economics	0.14	0.13	0.14
Major Description Engineering	0.06	0.06	0.06
Major Description None	0.20	0.25	0.22
Observations	48,719	39,947	88,666

NOTES: This table shows more information on applicants' characteristics, education histories, and work experience. The sample in Column 1 consists of all applicants who applied to a position during our training period (2016 and 2017). Column 2 consists of applicants who applied during the analysis period (2018 to Q1 2019). Column 3 presents summary statistics for the full pooled sample. All data come from the firm's application and hiring records.

TABLE A.2: CORRELATIONS BETWEEN ALGORITHM SCORES AND HIRING LIKELIHOOD

	Hired			Offered		
	(1)	(2)	(3)	(4)	(5)	(6)
Human	-0.0652**			-0.125***		
	(0.0280)			(0.0363)		
SL Hired		0.171***				
		(0.0267)				
UCB Hired			0.205***			
			(0.0261)			
SL Offered					0.284***	
					(0.0294)	
UCB Offered						0.349***
						(0.0311)
Observations	2275	2275	2275	2275	2275	2275
Mean of Hired: .102						
Mean of Offered: .189						

NOTES: This table presents the results of regressing an indicator for being hired on the algorithm scores on the sample of interviewed applicants in the analysis period. Control variables include fixed effects for job family, application year-month, and seniority level. All data come from the firm's application and hiring records. Robust standard errors shown in parentheses.

Table A.3: Average Monotonicity Test

	(1) All	(2) White	(3) Black	(4) Asian	(5) Hispanic	(6) Female	(7) Male	(8) MA	(9) Ref.
Interviewer Leniency	0.08***	0.08***	0.06***	0.09***	0.04	0.08***	0.08***	0.08***	0.10***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	23803	7055	2150	13565	1097	8825	16575	16384	3245

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Notes: This table presents results of regressing an applicant's interview status on their jack-knifed screener leniency instrument, by subgroups defined by race, education and gender characteristics. Each column presents the coefficient on the interviewer leniency variable for that subgroup regression. All specifications include controls for job type, job seniority level, work location and application date. Standard errors are clustered at the screener level. We find that being assigned to a lenient screener is on average related to the propensity to get an interview and the benefit is similar across demographic groups.

TABLE A.4: CORRELATION OF PREFERENCES OF LENIENT AND STRICT SCREENERS

	Within-Person Correlation Strict and Lenient Selection Scores
All	0.688
White	0.836
Black	0.860
Asian	0.702
Hispanic	0.760
Female	0.703
Male	0.686
MA	0.686
Referral	0.689

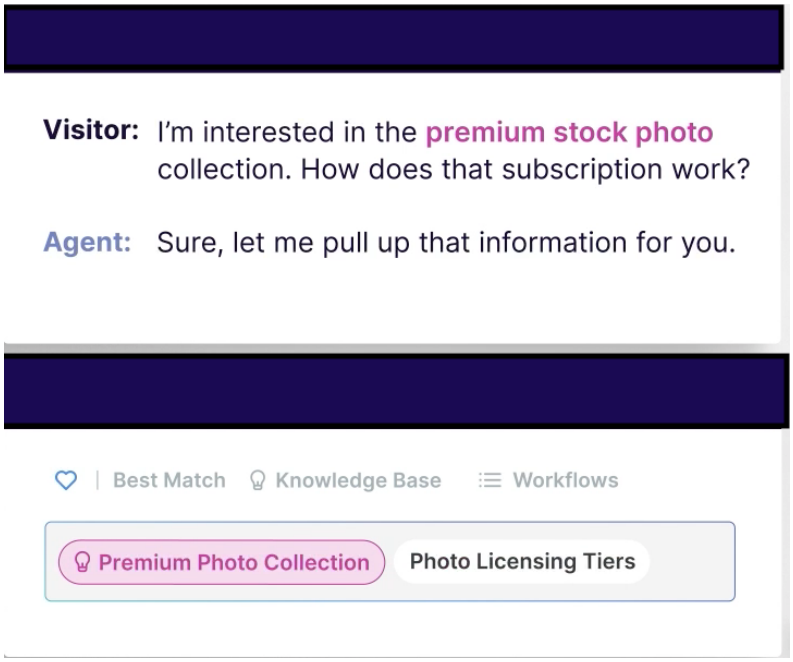
NOTES: This table presents the results of another test of monotonicity. Here, we predict two propensity scores for every applicant: that applicant's likelihood of being selected by a strict screener and that applicant's likelihood of being selected by a lenient screener. We then examine the correlation between these two scores across subgroups in order to ask whether the preferences of lenient and strict screeners are correlated. To do this, we randomly split our sample into a train and test set for applicants assigned to strict screeners and lenient screeners (above and below the median jack-knifed leniency). We train a regularized logit model that predicts interview propensity for the set of strict screeners and a second model on lenient screeners. During our testing period, we generate an out-of-sample predicted probability of interview on all applicants using the strict interviewer model and the lenient screener model. We find that the correlation between the predicted probability of interview under the lenient screener and the strict screeners is positive across race, gender, and education groups.

Appendix C

Generative AI at Work Appendix Materials

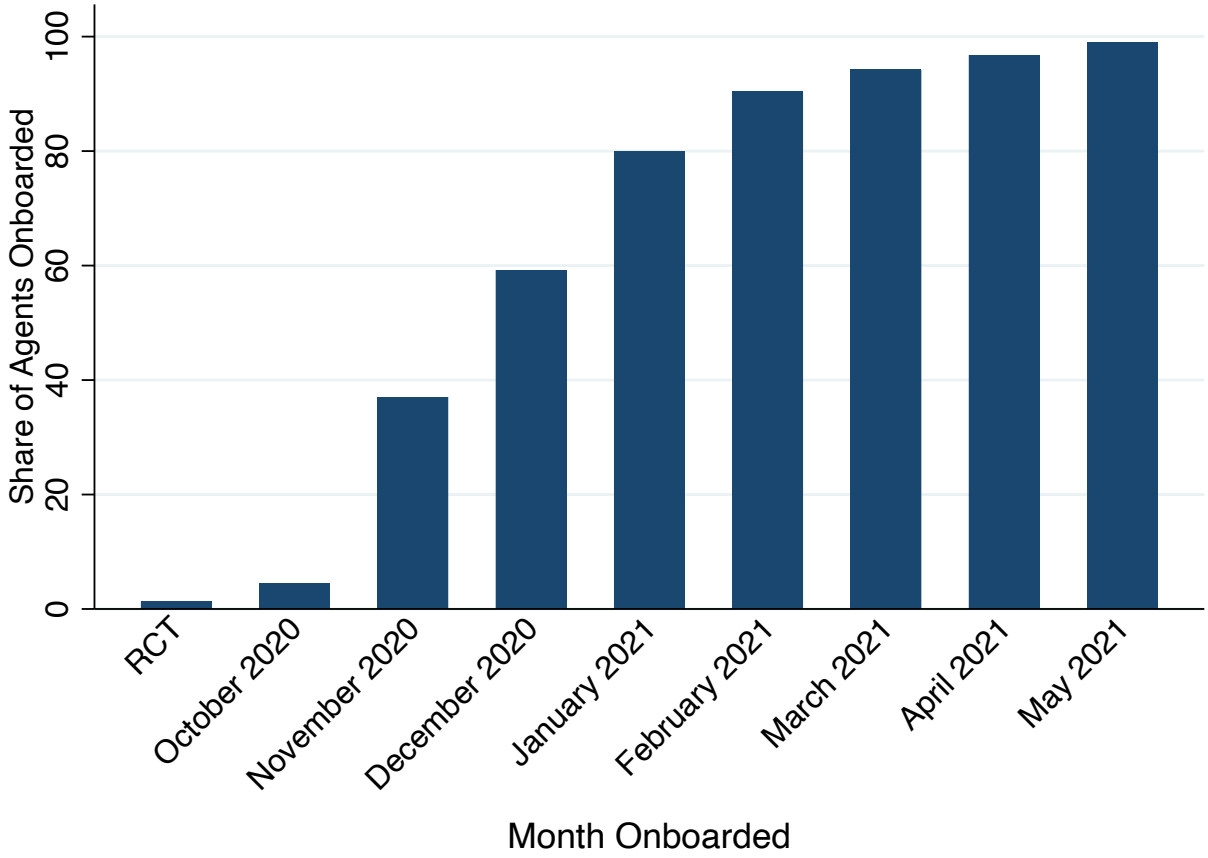
FIGURE A.1: SAMPLE AI TECHNICAL SUGGESTION

A. SAMPLE AI-GENERATED TECHNICAL LINK



NOTES: This figure shows a sample technical documentation suggestions made the by AI. Our data firm has an extensive set of documentation for their technical support agents, known as the knowledge base, which is like an internal company Wikipedia for product and process information. The AI will attempt to surface the most helpful technical documentation page when triggered to do so during a customer interaction. These links are only visible to the agent and agents must review to see if the resource is helpful. Workers can choose to read the suggested technical documentation or ignore the recommendation.

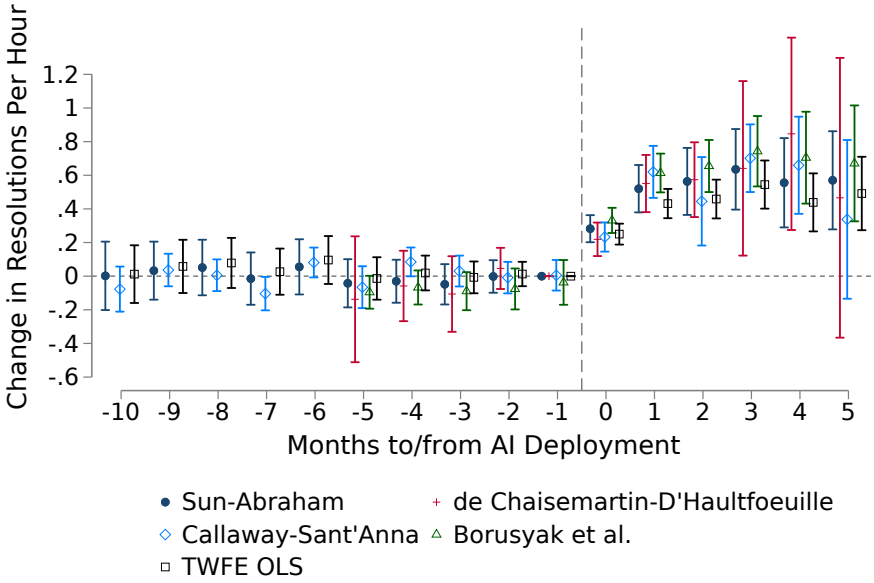
FIGURE A.2: DEPLOYMENT TIMELINE



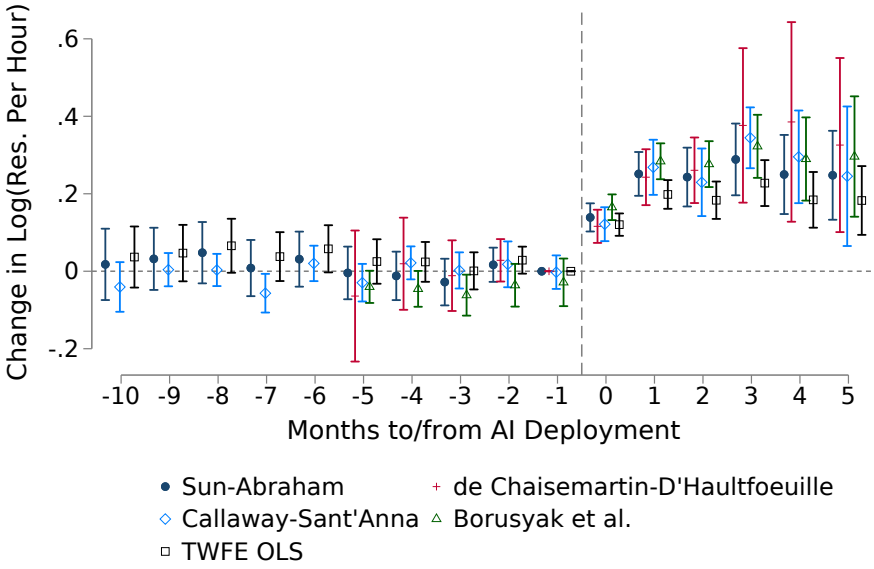
NOTES: This figure shows the share of agents deployed onto the AI system over the study period. Agents are deployed onto the AI system after a training session. The firm ran a small randomized control trial in August and September of 2020. All data are from the firm's internal software systems.

FIGURE A.3: EVENT STUDIES, RESOLUTIONS PER HOUR

A. RESOLUTIONS PER HOUR



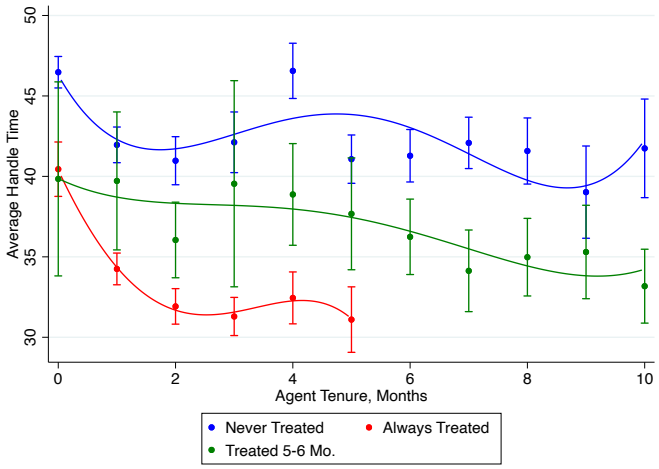
B. LOG(RESOLUTIONS PER HOUR)



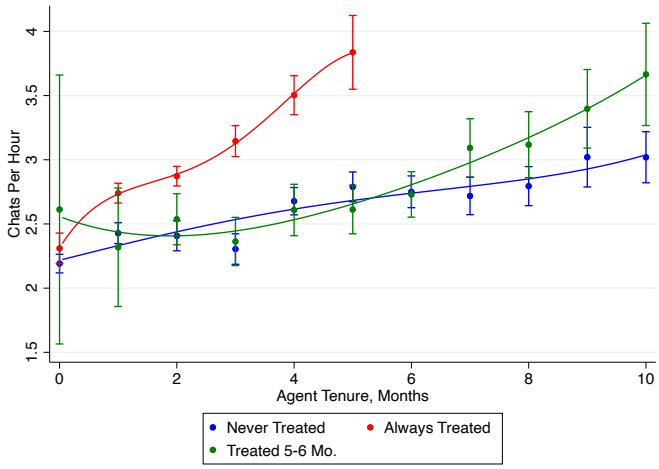
NOTES: This table presents the effect of AI model deployment on our main productivity outcome, resolutions per hour, using a variety of robust dynamic difference-in-differences estimators introduced in Borusyak, Jaravel and Spiess (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021) and a standard two-way fixed effects regression model. All regressions include agent level, chat-year fixed effects and controls for agent tenure. Standard errors are clustered at the agent level. Because of the number of post-treatment periods and high turnover of agents in our sample, we can only estimate five months of preperiod using Borusyak, Jaravel and Spiess (2022) and de Chaisemartin and D'Haultfœuille (2020).

Figure A.4: Experience Curves by Deployment Cohort, Additional Outcomes

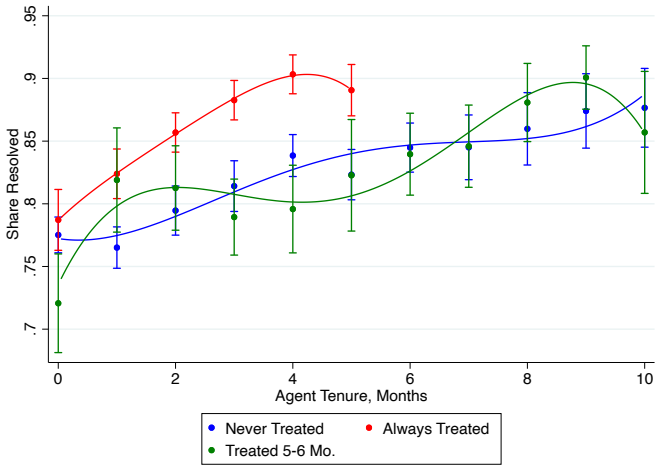
A. Average Handle Time



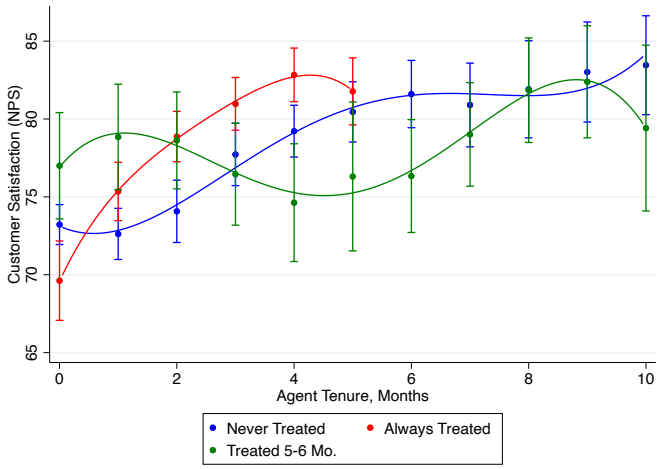
B. Chats Per Hour



C. Resolution Rate

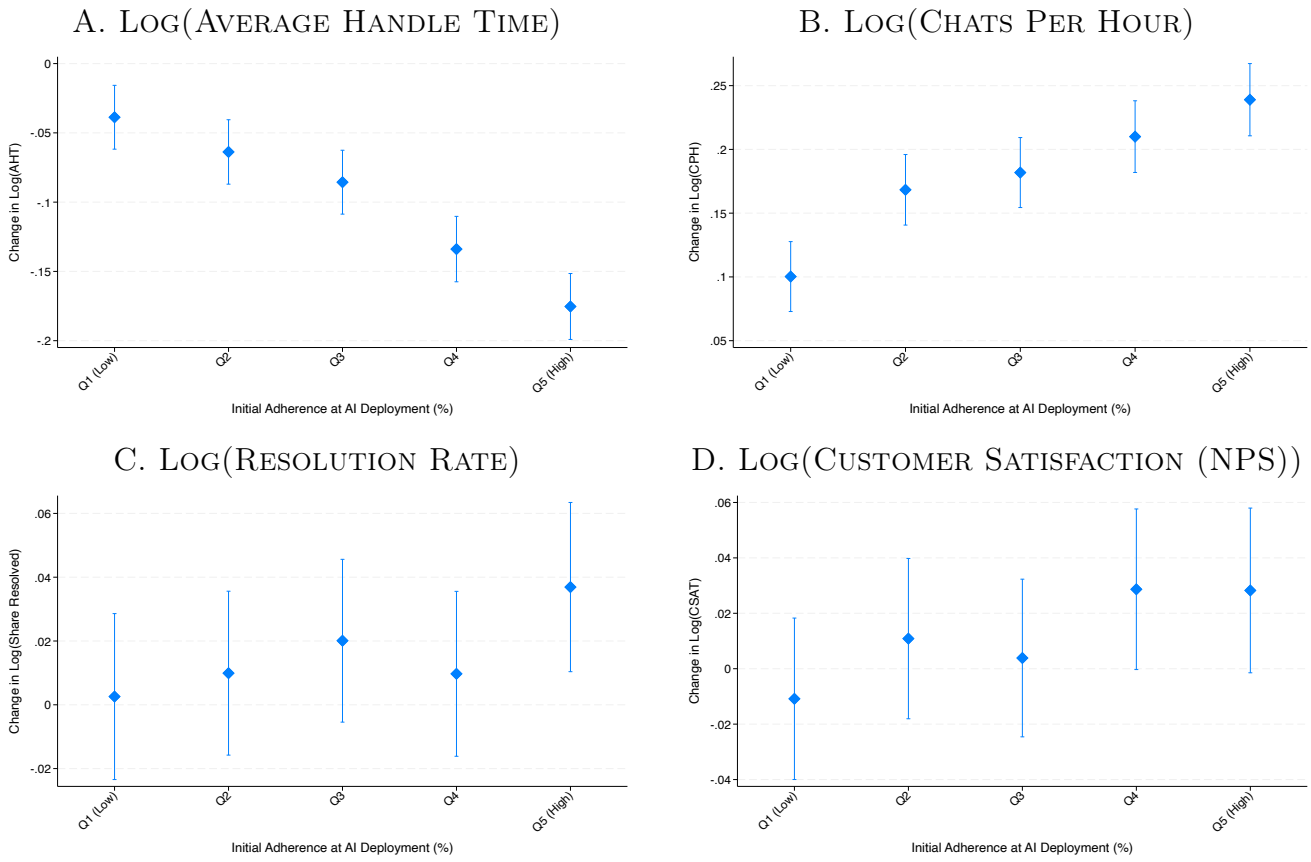


D. Customer Satisfaction



Notes: These figures plot the experience curves of three groups of agents over their tenure, the x-axis, against five measures of productivity and performance. The red lines plot the performance of always-treated agents, those who are start work in their first month with the AI and always have access to the AI suggestions. The blue line plots agents who are never treated. The green line plots agents who spend their first four months of work without the AI model, and gain access to the AI during their fifth month on the job. All panels include 95% confidence intervals.

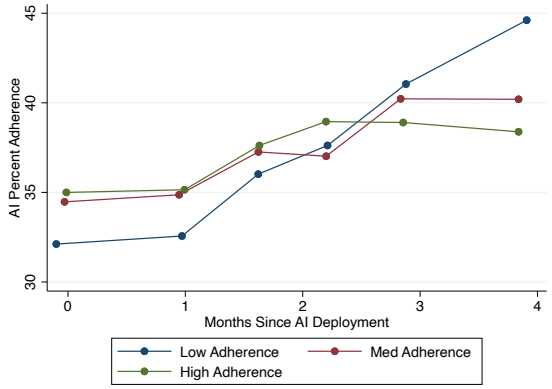
Figure A.5: Heterogeneity of AI Impact by Initial AI Adherence, Additional Outcomes



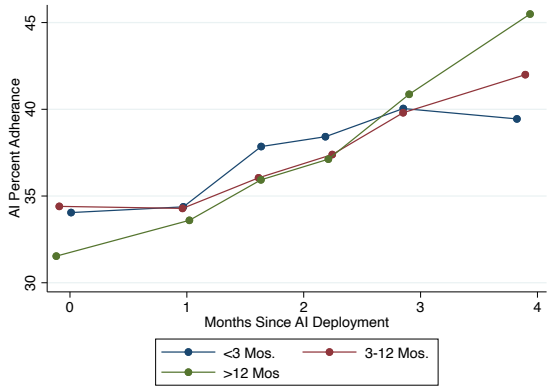
Notes: These figures plot the impact of AI model deployment on additional measures of performance by quintile of initial adherence, the share of AI recommendations followed in the first month of treatment. Panel A plots the average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of chats an agent can handle per hour (including working on multiple chats simultaneously). Panel C plots the resolution rate, the share of chats successfully resolved, and Panel D plots NPS, or net promoter score, is an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, and company fixed effects and controls for agent tenure. All data come from the firm's internal software systems.

Figure A.6: Within-agent AI Adherence over Time

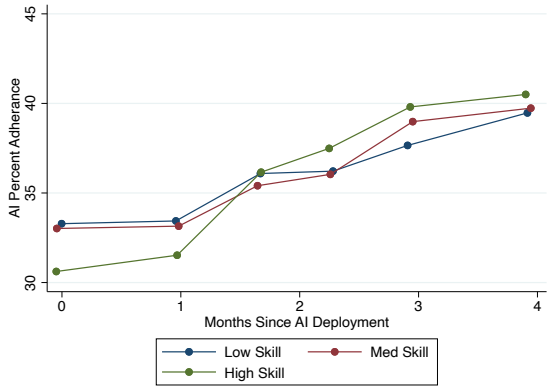
A. By Adherence at AI Model Deployment



B. By Agent Tenure at AI Model Deployment

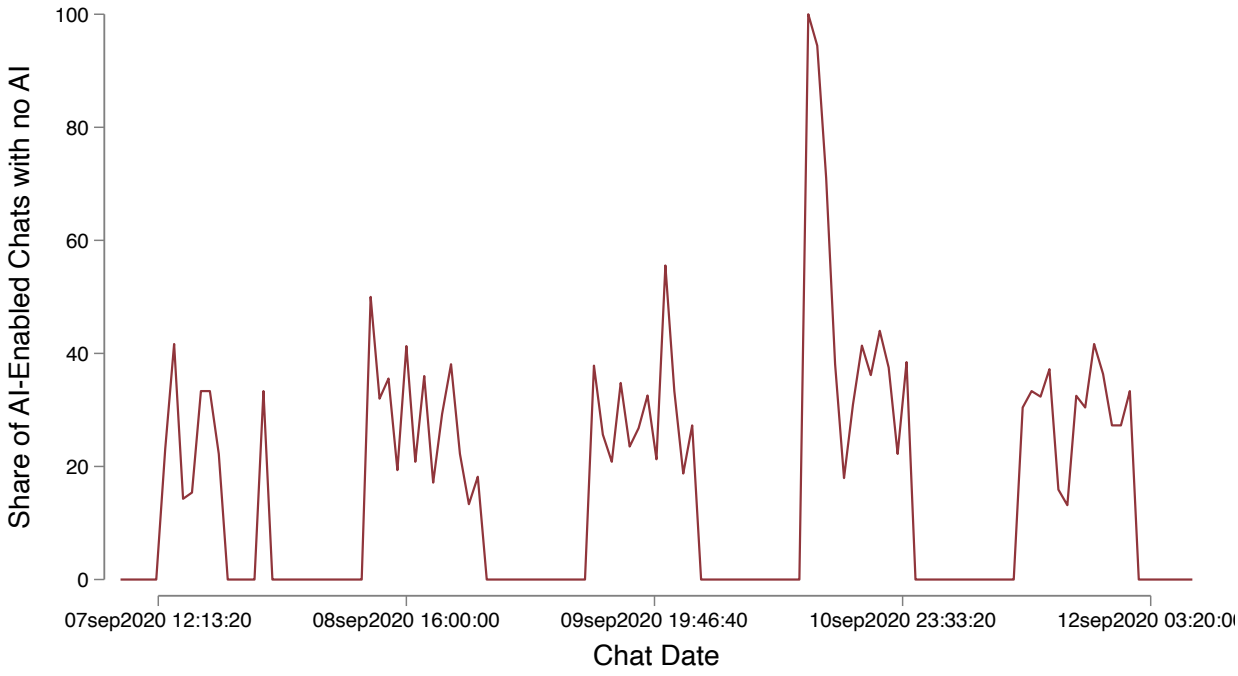


C. By Agent Skill at AI Model Deployment



Notes: This figure plots the residualized percentage of AI suggestions followed by agents as a function of the number of months each agent has had access to the AI model, after controlling for agent level fixed effects. In Panel A, we divide agents into terciles based on their adherence to AI suggestions in the first month. In Panel B, we divide agents into groups based on their tenure at the firm at the time of AI model deployment. In Panel C, we divide workers into terciles of pre-deployment productivity as defined by our skill index. All data come from the firm's internal software systems.

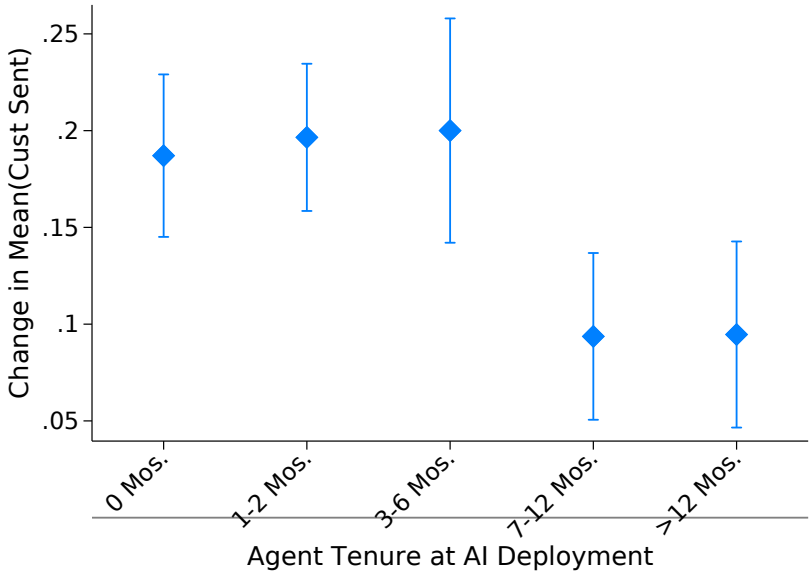
FIGURE A.7: SAMPLE AI OUTAGE



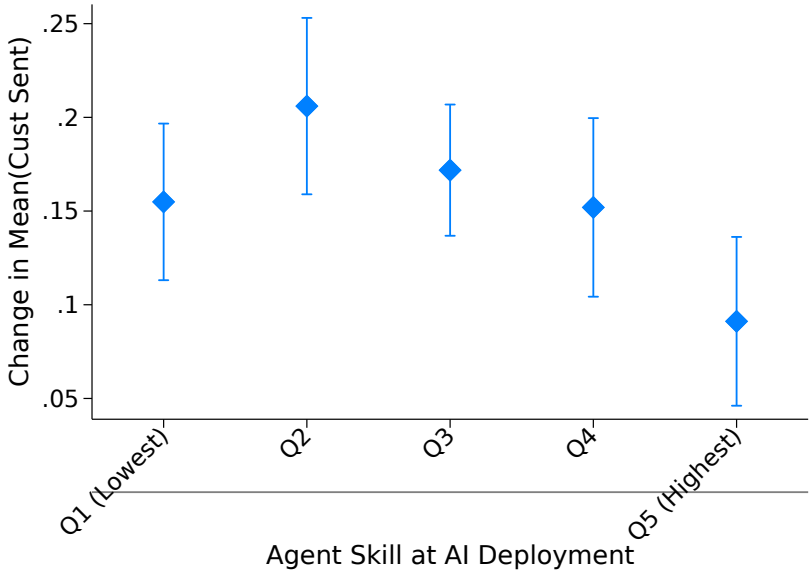
NOTES: This figure plots the share of post-treatment chats with no AI suggestions during a period of a documented software outage.

FIGURE A.8: HETEROGENEITY IN CUSTOMER SENTIMENT

A. BY TENURE AT AI MODEL DEPLOYMENT



B. BY PRODUCTIVITY AT AI MODEL DEPLOYMENT

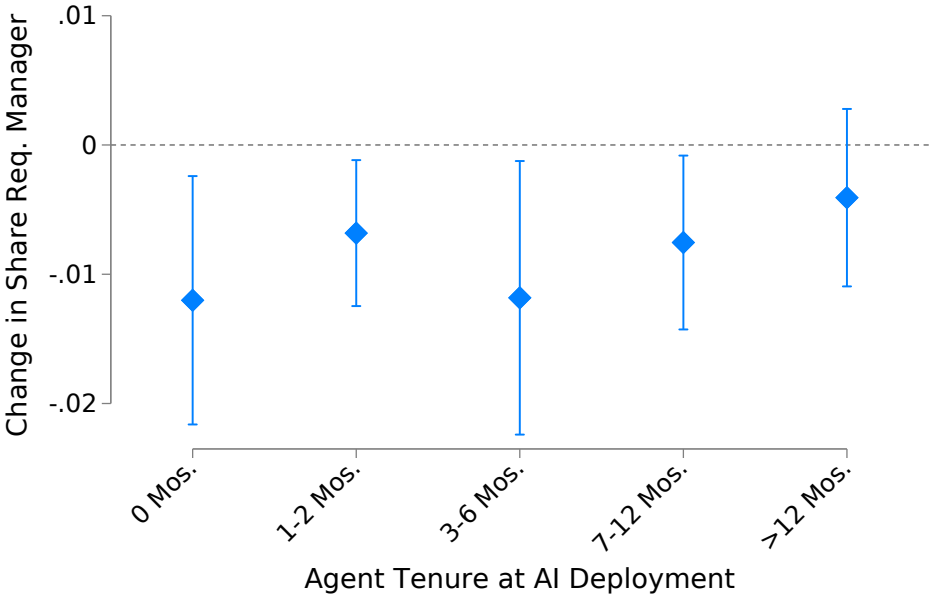


NOTES: Each panel of this figure plots the impact of AI model deployment on the mean sentiment per conversation. Sentiment refers to the emotion or attitude expressed in the text of the customer chat and ranges from -1 to 1 where -1 indicates very negative sentiment and 1 indicates very positive sentiment. Panel A plots the effects of AI model deployment on customer sentiment by agent tenure when AI deployed and Panel B plots the impacts by agent ex-ante productivity. All data come from the firm's internal software systems. Average sentiment is measured using SiEBERT, a fine-tuned checkpoint of a RoBERTA, an english language transformer model.

FIGURE A.9: ESCALATION, HETEROGENEITY BY WORKER TENURE AND SKILL

A. MANAGER ASSISTANCE, BY PRE-AI SKILL

B. MANAGER ASSISTANCE, BY PRE-AI TENURE



NOTES: Panels A and B show the effects of AI on customer requests for manager assistance, by pre-AI agent skill and in by pre-AI agent tenure. All robust standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

TABLE A.1: MAIN EFFECTS: PRODUCTIVITY (LOG(RESOLUTIONS PER HOUR)), ALTERNATIVE DIFFERENCE-IN-DIFFERENCE ESTIMATORS

	Point Estimate	Standard Error	Lower Bound 95% Confidence Interval	Upper Bound 95% Confidence Interval
TWFE-OLS	0.137	0.014	0.108	0.165
Borusyak-Jaravel-Spiess	0.257	0.028	0.203	0.311
Callaway-Sant'Anna	0.239	0.025	0.189	0.289
DeChaisemartin-D'Haultfoeuille	0.116	0.021	0.075	0.156
Sun-Abraham	0.237	0.037	0.165	0.308

NOTES: This table shows the impact of AI model deployment on the log of our main productivity outcome, resolutions per hour, using robust difference-in-differences estimators introduced in Borusyak, Jaravel and Spiess (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021). All regressions include agent level, chat-year fixed effects and controls for agent tenure. The standard errors are clustered at the agent level.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of econometrics*, 113(2): 231–263.

Abaluck, Jason, Leila Agha, David C. Chan Jr, Daniel Singer, and Diana Zhu. 2020. "Fixing Misallocation with Guidelines: Awareness vs. Adherence." National Bureau of Economic Research.

Acemoglu, Daron, and David Autor. 2011. "Skills, tasks and technologies: Implications for employment and earnings." In *Handbook of labor economics*. Vol. 4, 1043–1171. Elsevier.

Acemoglu, Daron, and Pascual Restrepo. 2018. "Low-Skill and High-Skill Automation." *Journal of Human Capital*, 12(2): 204–232.

Acemoglu, Daron, and Pascual Restrepo. 2020. "Robots and Jobs: Evidence from US Labor Markets." *Journal of Political Economy*, 128(6): 2188–2244. _eprint: https://doi.org/10.1086/705716.

Acemoglu, Daron, Gary Anderson, David Beede, Catherine Buffington, Eric Childress, Emin Dinlersoz, Lucia Foster, Nathan Goldschlag, John Haltiwanger, Zachary Kroff, Pascual Restrepo, and Nikolas Zolas. 2022. "Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey."

Acemoglu, Daron, Philippe Aghion, Claire Lelarge, John Van Reenen, and Fabrizio Zilibotti. 2007. "Technology, Information, and the Decentralization of the Firm*." *The Quarterly Journal of Economics*, 122(4): 1759–1799. _eprint: https://academic.oup.com/qje/article-pdf/122/4/1759/5234557/122-4-1759.pdf.

Agan, Amanda, and Sonja Starr. 2018. "Ban the box, criminal records, and racial discrimination: A field experiment." *The Quarterly Journal of Economics*, 133(1): 191–235.

Aggarwal, Nidhi, and Susan Thomas. 2014. "The causal impact of algorithmic trading on market quality."

Agrawal, Shipra, and Navin Goyal. 2013. "Further optimal regret bounds for thompson sampling." 99–107.

Akerman, Anders, Ingvil Gaarder, and Magne Mogstad. 2015. "The Skill Complementarity of Broadband Internet *." *The Quarterly Journal of Economics*, 130(4): 1781–1824. _eprint: https://academic.oup.com/qje/article-pdf/130/4/1781/30637431/qjv028.pdf.

Alston, Mackenzie. 2019. "The (Perceived) Cost of Being Female: An Experimental Investigation of Strategic Responses to Discrimination." *Working paper*.

American Homes 4 Rent. 2013. "Form S-11." *https://www.sec.gov/Archives/edgar/data/1562401/000119312513247145/d547003ds11.htm*.

American Homes 4 Rent. 2018. "10K-2017-Q4.Pdf." *https://s29.q4cdn.com/671712101/files/doc_financial/quarterly_results/2017/q4/10K-2017-Q4.pdf*.

Amherst. 2016. "U.S. Single Family Rental - An Emerging Institutional Asset Class." *https://www.amherst.com/insights/u-s-single-family-rental-an-emerging-institutional-asset-class/*.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association*, 91(434): 444–455.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics*, qjy012.

Arnold, David, Will S Dobbie, and Peter Hull. 2020. "Measuring Racial Discrimination in Algorithms." National Bureau of Economic Research Working Paper 28222.

Åslund, Olof, and Oskar Nordströum Skans. 2012. "Do anonymous job application procedures level the playing field?" *ILR Review*, 65(1): 82–107.

Athey, Susan, and Scott Stern. 1998*a*. "An Empirical Framework for Testing Theories About Complimentarity in Organizational Design." National Bureau of Economic Research Working Paper 6600. Series: Working Paper Series.

Athey, Susan, and Scott Stern. 1998*b*. "An Empirical Framework for Testing Theories About Complimentarity in Organizational Design." National Bureau of Economic Research Working Paper 6600.

Athey, Susan, and Scott Stern. 2002. "The Impact of Information Technology on Emergency Health Care Outcomes." *RAND Journal of Economics*, 33(3): 399–432.

Athey, Susan, Christopher Avery, and Peter Zemsky. 2000. "Mentoring and Diversity." *American Economic Review*, 90(4): 765–786.

Athey, Susan, Joshua Gans, Scott Schaefer, and Scott Stern. 1994. "The Allocation of Decisions in Organizations." *Stanford Graduate School of Business*.

Auer, Peter. 2002. "Using Confidence Bounds for Exploitation-Exploration Trade-offs." *Journal of Machine Learning Research*, 3(Nov): 397–422.

Autor, David. 2014. "Polanyi's Paradox and the Shape of Employment Growth." National Bureau of Economic Research Working Paper w20485.

Autor, David H, and David Scarborough. 2008. "Does job testing harm minority workers? Evidence from retail establishments." *The Quarterly Journal of Economics*, 123(1): 219–277.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics*, 118(4): 1279–1333.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?*." *The Quarterly Jour-*

nal of Economics, 113(4): 1169–1213. _eprint: https://academic.oup.com/qje/article-pdf/113/4/1169/5406877/113-4-1169.pdf.

Babcock, Frederick M. 1932. *The Valuation of Real Estate.* New York:McGraw Hill Book Company.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson. 2022. "Artificial Intelligence, Firm Growth, and Product Innovation."

Bagues, Manuel, and Chris Roth. 2021. "Interregional Contact and National Identity." *CEPR Working Paper 15576.*

Bagues, Manuel F., and Berta Esteve-Volart. 2010. "Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment." *Review of Economic Studies*, 77(4): 1301 – 1328.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate."

Bakalar, Chloé, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. 2021. "Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems."

Baker, George P., and Thomas N. Hubbard. 2003. "Make Versus Buy in Trucking: Asset Ownership, Job Design, and Information." *American Economic Review*, 93(3): 551–572.

Barocas, Solon, and Andrew D Selbst. n.d.. "Big Data's Disparate Impact." *Calif. L. Rev.. California Law Review*, 104(IR): 671.

Bartel, Ann, Casey Ichniowski, and Kathryn Shaw. 2007. "How Does Information Technology Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement, and Worker Skills*." *The Quarterly Journal of Economics*, 122(4): 1721–1758.

Bartos, Vojtech, Michal Bauer, Julie Chytilova, and Filip Matejka. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review*, 106(6): 1437–75.

Bastani, Hamsa, and Mohsen Bayati. 2019. "Online Decision-Making with High-Dimensional Covariates." 58.

Bastani, Hamsa, Mohsen Bayati, and Khashayar Khosravi. 2019. "Mostly Exploration-Free Algorithms for Contextual Bandits." *arXiv:1704.09011 [cs, stat]*. arXiv: 1704.09011.

Bayer, Patrick, Marcus Casey, Fernando Ferreira, and Robert McMillan. 2017. "Racial and Ethnic Price Differentials in the Housing Market." *Journal of Urban Economics*, 102: 91–105.

Bechavod, Yahav, Katrina Ligett, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. 2020. "Equal Opportunity in Online Classification with Partial Feedback."

Becker, Gary S. (Gary Stanley). 1957. *The Economics of Discrimination. Studies in Economics of the Economics Research Center of the University of Chicago*, Chicago:University of Chicago Press.

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon. 2015. "Unintended effects of anonymous resumes." *American Economic Journal: Applied Economics*, 7(3): 1–27.

Benjamin, Daniel J. 2019. "Chapter 2 - Errors in probabilistic reasoning and judgment biases." In *Handbook of Behavioral Economics - Foundations and Applications 2*. Vol. 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 69–186. North-Holland.

Benson, Alan, Danielle Li, and Kelly Shue. 2019. "Promotions and the Peter Principle." *The Quarterly Journal of Economics*, 134(4): 2085–2134.

Benson, Alan, Danielle Li, and Kelly Shue. 2021. "Potential and the Gender Promotion Gap." *mimeo*.

Berg, Jeff, Avinash Das, Vinay Gupta, and Paul Kline. 2018. "Smarter call-center coaching for the digital world." McKinsey & Company.

Bertrand, Marianne, and Esther Duflo. 2017. "Field Experiments on Discrimination." In *Handbook of Field Experiments*. Vol. 1 of *Handbook of Economic Field Experiments*, , ed. Abhijit Vinayak Banerjee and Esther Duflo, 309–393. North-Holland.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit discrimination." *American Economic Review*, 95(2): 94–98.

Bhutta, Neil, Andrew C. Chang, Lisa J. Dettling, and Joanne W. Hsu with assistance from Julia Hewitt. 2020. "Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances." *FEDS Notes*.

Blattner, Laura, and Scott Nelson. 2021. "How Costly Is Noise? Data and Disparities in Consumer Credit." Comment: 86 pages, 17 figures.

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature*, 55(3): 789–865.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen. 2014. "The Distinct Effects of Information Technology and Communication Technology on Firm Organization." *Management Science*, 60(12): 2859–2885.

BLS. 2019. "Industries with the largest wage and salary employment growth and declines."

Bogen, Miranda, and Aaron Rieke. 2018*a*. "Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias."

Bogen, Miranda, and Aaron Rieke. 2018*b*. "Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias." *Upturn*.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review*, 109(10): 3395–3436.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope. 2019. "Inaccurate Statistical Discrimination: An Identification Problem." National Bureau of Economic Research Working Paper 25935.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2022. "Revisiting Event Study Designs: Robust and Efficient Estimation."

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. 2002. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence." *The Quarterly Journal of Economics*, 117(1): 339–376.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. "Language Models are Few-Shot Learners." arXiv:2005.14165 [cs].

Brown, Zach Y., and Alexander MacKay. 2023. "Competition in Pricing Algorithms." *American Economic Journal: Microeconomics*, 15(2): 109–56.

Brynjolfsson, Erik, and Tom Mitchell. 2017. "What Can Machine Learning, Do? Workforce Implications." *Science*, 358: 1530–1534.

Brynjolfsson, Erik, Lindsey Raymond, and Danielle Li. 2023. "Generative AI at Work." *NBER Working Paper No. 31161.*

Bubeck, Sebastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. "Sparks of artificial general intelligence: Early experiments with gpt-4." *arXiv preprint arXiv:2303.12712.*

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru. 2022. "Why Is Intermediating Houses so Difficult? Evidence from iBuyers."

Buesing, Eric, Vinay Gupta, Sarah Higgins, and Raelyn Jacobson. 2020. "Customer care: The future talent factory." McKinsey & Company.

Busse, Meghan R, Devin G Pope, Jaren C Pope, and Jorge Silva-Risso. 2012. "Projection Bias in the Car and Housing Markets." National Bureau of Economic Research Working Paper 18212.

Calder-Wang, Sophie, and Gi Heung Kim. 2023. "Coordinated vs Efficient Prices: The Impact of Algorithmic Pricing on Multifamily Rental Markets."

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. 2020. "Artificial Intelligence, Algorithmic Pricing, and Collusion." *American Economic Review*, 110(10): 3267–97.

Calvino, Flavio, and Luca Fontanelli. 2023. "A Portrait of AI Adopters across Countries: Firm Characteristics, Assets' Complementarities and Productivity." OECD, Paris.

Casey, Joan A., Frosch Rachel Morello, Daniel J. Mennitt, Kurt Fristrup, Elizabeth L. Ogburn, and Peter James. 2017. "Race/Ethnicity, Socioeconomic Status,

Residential Segregation, and Spatial Variation in Noise Exposure in the Contiguous United States." *Environmental Health Perspectives*, 125(7): 077017.

Castilla, Emilio. 2011. "Bringing Managers Back In." *American Sociological Review*, 76: 667–694.

Castilla, Emilio J. 2008. "Gender, Race, and Meritocracy in Organizational Careers." *American Journal of Sociology*, 113(6): 1479–1526.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2018. "Manipulation Testing Based on Density Discontinuity." *The Stata Journal: Promoting communications on statistics and Stata*, 18(1): 234–261.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2019. "Simple Local Polynomial Density Estimators."

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs*." *The Quarterly Journal of Economics*, 134(3): 1405–1454.

Census. 2023. "Census Tabulation Detail: Tenure by Units in Structure." *https://censusreporter.org/tables/B25032/.*

Chaboud, Alain P., Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega. 2014. "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market." *The Journal of Finance*, 69(5): 2045–2084.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost." ACM.

Choi, Jonathan H., and Daniel Schwarcz. 2023. "AI Assistance in Legal Analysis: An Empirical Study."

Choi, Jung Hyun, Caitlin Young, Alanna McCargo, Michael Neal, and Laurie Goodman. 2019. "Explaining the Black-White Homeownership Gap." *The Urban Institute.*

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions." 134–148. PMLR.

Christophers, Brett. 2023. "How and Why U.S. Single-Family Housing Became an Investor Asset Class." *Journal of Urban History*, 49(2): 430–449.

Chui, Michael, Bryce Hall, Alex Singla, and Alex Sukharevsky. 2021. "Global survey: The state of AI in 2021." McKinsey & Company.

Clark, Robert, Stephanie Assad, Daniel Ershov, and Lei Xu. 2023. "Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market." *Journal of Political Economy*, 0(ja): null.

Consumer Financial Protection Bureau. 2023. "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity." *https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/.*

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv:1808.00023 [cs]*. arXiv: 1808.00023.

Cowgill, Bo. 2020. "Bias and productivity in humans and algorithms: Theory and evidence from resume screening." *Columbia Business School, Columbia University*, 29.

Cowgill, Bo, and Catherine E Tucker. 2019. "Economics, fairness and algorithmic bias." *preparation for: Journal of Economic Perspectives*.

Craigie, Terry-Ann. 2020. "Ban the Box, Convictions, and Public Employment." *Economic Inquiry*, 58(1): 425–445.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 1999. "The Rise and Decline of the American Ghetto." *Journal of Political Economy*, 107(3): 455–506.

Dawes, Robyn M. 1971. "A case study of graduate admissions: Application of three principles of human decision making." *American Psychologist*, 26(2): 180–188.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science*, 243(4899): 1668–1674.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–96.

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Quarterly Journal of Economics*, 132(4): 1593–1640.

Derenoncourt, Ellora, Chi Hyun Kim, Moritz Kuhn, and Moritz Schularick. 2022. "Wealth of Two Nations: The U.S. Racial Wealth Gap, 1860-2020." National Bureau of Economic Research Working Paper 30101.

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens. 2018*a*. "Balanced Linear Contextual Bandits." *arXiv:1812.06227 [cs, stat]*. arXiv: 1812.06227.

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens. 2018*b*. "Estimation Considerations in Contextual Bandits." *arXiv:1711.07077 [cs, econ, stat]*. arXiv: 1711.07077.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review*, 108(2): 201–40.

Doleac, Jennifer L, and Benjamin Hansen. 2020. "The unintended consequences of "ban the box": Statistical discrimination and employment outcomes when criminal histories are hidden." *Journal of Labor Economics*, 38(2): 321–374.

Dunn, Andrew, Diana Inkpen, and Răzvan Andonie. 2021. "Context-Sensitive Visualization of Deep Learning Natural Language Processing Models."

Eagly, A. H., and S. J. Karau. 2002. "Role congruity theory of prejudice toward female leaders." *Psychological Review*, 109(3): 573–598.

Economist, The. 2021. "A Whodunnit on Zillow." *The Economist*.

Einav, Liran, Mark Jenkins, and Jonathan Levin. 2013. "The impact of credit scoring on consumer lending." *The RAND Journal of Economics*, 44(2): 249–274.

Elliott, Marc N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities." *Health Services and Outcomes Research Methodology*, 9(2): 69–83.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." arXiv:2303.10130 [cs, econ, q-fin].

Elster, Yael, and Noam Zussman. 2022. "Minorities and Property Values: Evidence from Residential Buildings in Israel." *Journal of Urban Economics*, 103525.

Erickson, Nick, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data."

Felten, Edward W., Manav Raj, and Robert Seamans. 2023. "Occupational Heterogeneity in Exposure to Generative AI."

Fernandez, Roberto M., Emilio J. Castilla, and Paul Moore. 2000. "Social Capital at Work: Networks and Employment at a Phone Center." *American Journal of Sociology*, 105(5): 1288–1356.

Fields, Desiree. 2018. "Constructing a New Asset Class: Property-led Financial Accumulation after the Crisis." *Economic Geography*, 94(2): 118–140.

Fields, Desiree. 2022. "Automated Landlord: Digital Technologies and Post-Crisis Financial Accumulation." *Environment and Planning A Economy and Space*, 54(1): 160–181.

Fischer, Christine, Karoline Kuchenbäcker, Christoph Engel, Silke Zachariae, Kerstin Rhiem, Alfons Meindl, Nils Rahner, Nicola Dikow, Hansjörg Plendl, Irmgard Debatin, et al. 2013. "Evaluating the performance of the breast cancer genetic risk models BOADICEA, IBIS, BRCAPRO and Claus for predicting BRCA1/2 mutation carrier probabilities: a study based on 7352 families from the German Hereditary Breast and Ovarian Cancer Consortium." *Journal of medical genetics*, 50(6): 360–367.

Fischhoff, B., P. Slovic, and S. Lichtenstein. 1977. "Knowing with certainty: The appropriateness of extreme confidence." *Journal of Experimental Psychology: Human Perception and Performance*, 3(4): 1124–1131.

Francke, Marc, Lianne Hans, Matthijs Korevaar, and Sjoerd van Bekkum. 2023. "Buy-to-Live vs. Buy-to-Let: The Impact of Real Estate Investors on Housing Costs and Neighborhoods."

Frandsen, Brigham R, Lars J Lefgren, and Emily C Leslie. 2019. "Judging Judge Fixed Effects." National Bureau of Economic Research Working Paper 25528.

Frankel, Alex. 2021. "Selecting Applicants." *Econometrica*, 89(2): 615–645.

Freddie Mac Economic & Housing Research. 2018. "Single-Family Rental: An Evolving Market." Federal Home Loan Mortgage Corporation.

Freddie Mac Economic & Housing Research. 2021. "Racial and Ethnic Valuation Gaps in Home Purchase Appraisals." Federal Home Loan Mortgage Corporation.

Friedman, Sam, and Daniel Laurison. 2019. *The Class Ceiling: Why It Pays to Be Privileged.* University of Chicago Press.

Fuster, Andres, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *The Journal of Finance*, 77(1): 5–47.

Gabaix, Xavier. 2019. "Chapter 4 - Behavioral inattention." In *Handbook of Behavioral Economics - Foundations and Applications 2*. Vol. 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 261–343. North-Holland.

Garicano, Luis. 2000. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy*, 108(5): 874–904. Publisher: The University of Chicago Press.

Garicano, Luis, and Esteban Rossi-Hansberg. 2015. "Knowledge-Based Hierarchies: Using Organizations to Understand the Economy." *Annual Review of Economics*, 7(1): 1–30.

Gillis, Talia B, and Jann L Spiess. 2019. "Big Data and Discrimination." *The University of Chicago Law Review*, 29.

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating impartiality: The impact of "blind" auditions on female musicians." *American economic review*, 90(4): 715–741.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.

Google. n.d.. "AI vs. Machine Learning: How Do They Differ?"

Gretz, Whitney, and Raelyn Jacobson. 2018. "Boosting contact-center performance through employee engagement." McKinsey & Company.

Guren, Adam M, Alisdair McKay, Emi Nakamura, and Jón Steinsson. 2020. "Housing Wealth Effects: The Long View." *The Review of Economic Studies*, 88(2): 669–707.

Gurun, Umit G, Jiabin Wu, Steven Chong Xiao, and Serena Wenjing Xiao. 2023. "Do Wall Street Landlords Undermine Renters' Welfare?" *The Review of Financial Studies*, 36(1): 70–121.

Harris, Adam, and Maggie Yellen. 2023. "Human Decision-Making with Machine Prediction: Evidence from Predictive Maintenance in Trucking."

Harris, David. 1999. ""Property Values Drop When Blacks Move in, Because...": Racial and Socioeconomic Determinants of Neighborhood Desirability." *American Sociological Review*, 64(3): 461–479.

Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. "More than a Feeling: Accuracy and Application of Sentiment Analysis." *International Journal of Research in Marketing*, 40(1): 75–87.

Hastie, R., and R. M. Dawes. 2001*a*. *Rational choice in an uncertain world: The psychology of judgment and decision making.* Sage Publications.

Hastie, R., and R. M. Dawes. 2001*b*. *Rational choice in an uncertain world: The psychology of judgment and decision making.* Sage Publications.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld. 2011. "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance*, 66(1): 1–33.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 71(4): 1161–1189. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00442.

Hirshleifer, David, Yaron Levi, Ben Lourie, and Siew Hong Teoh. 2019. "Decision fatigue and heuristic analyst forecasts." *Journal of Financial Economics*, 133(1): 83–98.

Hochschild, Arlie Russell. 2019. *The managed heart: Commercialization of human feeling.* University of California press.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li. 2017. "Discretion in Hiring*." *The Quarterly Journal of Economics*, 133(2): 765–800.

Housman, Michael, and Dylan Minor. 2015. "Toxic Workers." Harvard Business School Working Paper 16-057.

Howell, Junia, and Elizabeth Korver-Glenn. 2018. "Neighborhoods, Race, and the Twenty-first-century Housing Appraisal Industry." *Sociology of Race and Ethnicity*, 4: 473 – 490.

Hugging Face. 2023. "sentence-transformers/all-MiniLM-L6-v2."

Invitation Homes. 2017. "2017-Annual-Report.Pdf." *https://s28.q4cdn.com/264003623/files-/doc_financials/2017/ar/2017-Annual-Report.pdf.*

Jackson, Summer. 2020. "Not Paying for Diversity: Repugnance and Failure to Choose Labor Market Platforms that Facilitate Hiring Racial Minorities into Technical Positions."

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment.* . First edition ed., New York:Little, Brown Spark.

Kamin, Debra. 2023. "Home Appraised With a Black Owner: $472,000. With a White Owner: $750,000. - The New York Times." *New York Times.*

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361.*

Kasy, Maximilian, and Rediet Abebe. 2020. "Fairness, equality, and power in algorithmic decision making." *Workshop on Participatory Approaches to Machine Learning, International Conference on Machine Learning.*

Katz, Lawrence F., and Kevin M. Murphy. 1992. "Changes in Relative Wages, 1963-1987: Supply and Demand Factors." *The Quarterly Journal of Economics*, 107(1): 35–78.

Kausel, Edgar E., Satoris S. Culbertson, and Hector P. Madrid. 2016. "Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring decisions." *Organizational Behavior and Human Decision Processes*, 137: 27–44.

Kaysen, Ronda. 2023. "New Legislation Proposes to Take Wall Street Out of the Housing Market." *The New York Times.*

Kermani, Amir, and Francis Wong. 2021*a*. "Racial Disparities in Housing Returns." National Bureau of Economic Research Working Paper 29306.

Kermani, Amir, and Francis Wong. 2021*b*. "Racial Disparities in Housing Returns."

Kim, Sunwoong. 2000. "Race and Home Price Appreciation in Urban Neighborhoods: Evidence from Milwaukee, Wisconsin." *The Review of Black Political Economy*, 28(2): 9–28.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017*a*. "Human Decisions and Machine Predictions*." *The Quarterly Journal of Economics*, 133(1): 237–293.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017*b*. "Human decisions and machine predictions*." *The Quarterly Journal of Economics*, 133(1): 237–293. tex.eprint: https://academic.oup.com/qje/article-pdf/133/1/237/30636517/qjx032.pdf.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018*a*. "Human decisions and machine predictions." *The quarterly journal of economics*, 133(1): 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018*b*. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis*, 10.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review*, 105(5): 491–495.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv:1609.05807 [cs, stat]*. arXiv: 1609.05807.

Kline, Patrick M, Evan Rose, and Christopher R Walters. 2022. "Systemic Discrimination Among Large US Employers." National Bureau of Economic Research Working Paper 29053.

Kling, Jeffrey R. 2006. "Incarceration length, employment, and earnings." *American Economic Review*, 96(3): 863–876.

Kolev, Julian, Yuly Fuentes-Medel, and Fiona Murray. 2019. "Is Blinded Review Enough? How Gendered Outcomes Arise Even Under Anonymous Evaluation." National Bureau of Economic Research Working Paper 25759.

Korinek, Anton. 2022. "How innovation affects labor markets: An impact assessment." Brookings Institution Working Paper.

Koroteev, M. V. 2021. "BERT: A Review of Applications in Natural Language Processing and Understanding."

Kuhnen, Camelia M., and Paul Oyer. 2016. "Exploration for Human Capital: Evidence from the MBA Labor Market." *Journal of Labor Economics*, 34(S2): S255–S286.

Kuhn, Peter J, and Lizi Yu. 2019. "How Costly is Turnover? Evidence from Retail." National Bureau of Economic Research.

Lai, Tze Leung, and Herbert Robbins. 1985. "Asymptotically efficient adaptive allocation rules." *Advances in applied mathematics*, 6(1): 4–22.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables." 275–284.

LaValle, Steven M, Michael S Branicky, and Stephen R Lindemann. 2004. "On the relationship between classical grid search and probabilistic roadmaps." *The International Journal of Robotics Research*, 23(7-8): 673–692.

Lazear, Edward. 1998. "Hiring Risky Workers." In *Internal Labour Markets, Incentives and Employment.* , ed. Tachibanaki T. Ohashi I. Palgrave Macmillan, London.

Lee, Don. 2015. "The Philippines has become the call-center capital of the world." *Los Angeles Times*. Section: Business.

Legg, Shane, Marcus Hutter, et al. 2007. "A collection of definitions of intelligence." *Frontiers in Artificial Intelligence and applications*, 157: 17.

Lei, Huitian, Ambuj Tewari, and Susan A. Murphy. 2017. "An Actor-Critic Contextual Bandit Algorithm for Personalized Mobile Health Interventions."

Lepage, Louis-Pierre. 2020*a*. "Endogenous Learning, Persistent Employer Biases, and Discrimination." *University of Michigan, mimeo.*

Lepage, Louis-Pierre. 2020*b*. "Experimental Evidence on Endogenous Belief Formation in Hiring and Discrimination." *University of Michigan, mimeo.*

Leslie, Emily, and Nolan G. Pope. 2017. "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments." *The Journal of Law and Economics*, 60(3): 529–557.

Lewis, Valerie A., Michael O. Emerson, and Stephen L. Klineberg. 2011. "Who We'll Live With: Neighborhood Racial Composition Preferences of Whites, Blacks and Latinos." *Social Forces*, 89(4): 1385–1407.

Li, Chun. 2020. "OpenAI's GPT-3 Language Model: A Technical Overview."

Li, Danielle, Lindsey Raymond, and Peter Bergman. 2020. "Hiring as Exploration." *NBER Working Paper No. 27736.*

Lilien, Jake. 2023. "Faulty Foundations: Mystery-Shopper Testing In Home Appraisals Exposes Racial Bias Undermining Black Wealth."

Li, Lihong, Wei Chu, John Langford, and Robert E. Schapire. 2010. "A Contextual-Bandit Approach to Personalized News Article Recommendation." *Proceedings of the 19th international conference on World wide web - WWW '10*, 661. arXiv: 1003.0146.

Li, Lihong, Yu Lu, and Dengyong Zhou. 2017. "Provably Optimal Algorithms for Generalized Linear Contextual Bandits." *ICML'17*, 2071–2080. JMLR.org.

Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models." arXiv:2304.01852 [cs].

Malone, Thomas. 2023. "Residential Real Estate: Largest US Asset Class but Not Biggest Economic Driver." *https://www.corelogic.com/intelligence/why-the-uss-largest-asset-class-residential-real-estate-does-not-substantially-contribute-to-the-economic-output/.*

McKinney, Scott Mayer, et. al. 2020. "International evaluation of an AI system for breast cancer screening." *Nature*, 577(7788): 89–94.

Meijer, Erik. 2018. "Behind every great deep learning framework is an even greater programming languages concept (keynote)." 1–1.

Mejova, Yelena. 2009. "Sentiment Analysis: An Overview." *University of Iowa, Computer Science Department.*

Mercer. 2020. "Global Talent Trends."

Merkley, Mr, and Adam Smith. 2023. "End Hedge Fund Control of American Homes Act of 2023."

Michaels, Guy, Ashwini Natraj, and John Van Reenen. 2014. "Has ICT Polarized Skill Demand? Evidence from Eleven Countries Over Twenty-Five Years." *The Review of Economics and Statistics*, 96(1): 60–77.

Miller, Conrad. 2017. "The Persistent Effect of Temporary Affirmative Action." *American Economic Journal: Applied Economics*, 9(3): 152–90.

Mills, James, Raven Molloy, and Rebecca Zarutskie. 2019. "Large-Scale Buy-to-Rent Investors in the Single-Family Housing Market: The Emergence of a New Asset Class." *Real Estate Economics*, 47(2): 399–430.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2022. "Managing Self-Confidence: Theory and Experimental Evidence." *Management Science*, 68(11): 7793–7817.

Mullainathan, Sendhil, and Ashesh Rambachan. 2023. "From Predictive Algorithms to Automatic Generation of Anomalies."

Mullainathan, Sendhil, and Ziad Obermeyer. 2019. "Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error." *NBER WP.*

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care*." *The Quarterly Journal of Economics*, 137(2): 679–727.

Neal, Michael, Laurie Goodman, and Caitlin Young. 2020. "Housing Supply Chartbook." *The Urban Institute.*

Nguyen, Nhan, and Sarah Nadi. 2022. "An Empirical Evaluation of GitHub Copilot's Code Suggestions." 1–5. ISSN: 2574-3864.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Available at SSRN 4375283.*

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine." *The New England journal of medicine*, 375(13): 1216–1219.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the

health of populations." *Science (New York, N.Y.)*, 366(6464): 447–453. tex.eprint: https://www.science.org/doi/pdf/10.1126/science.aax2342.

OECD. 2023. *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market.* Paris:Organisation for Economic Co-operation and Development.

OpenAI. 2023. "GPT-4 Technical Report." OpenAI.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. "Training language models to follow instructions with human feedback." arXiv:2203.02155 [cs].

Pager, Devah, and Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology*, 34(1): 181–209. PMID: 20689680.

Paluck, Elizabeth Levy, and Donald P. Green. 2009. "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice." *Annual Review of Psychology*, 60(1): 339–367. PMID: 18851685.

Patel, Dylan, and Gerald Wong. 2023. "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE."

Peng, Baolin, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023*a*. "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback."

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023*b*. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot."

Perry, Andre, Jonathan Rothwell, and David Harshbarger. 2018. "The Devaluation of Assets in Black Neighborhoods."

Perry, Jonathan Rothwell and Andre M. 2021. "Biased Appraisals and the Devaluation of Housing in Black Neighborhoods."

Pew Research Center. 2018. Pew Research Center.

Pinto, Edward, and Tobias Peter. 2021. "AEI Housing Center Response to Perry and Rothwell (2021)." *American Enterprise Institute Housing Center.*

Polanyi, Michael. 1966. *The Tacit Dimension.* Chicago, IL:University of Chicago Press.

Prendergast, Canice, and Robert Topel. 1993. "Discretion and bias in performance evaluation." *European Economic Review*, 37(2-3): 355–365.

Quadlin, Natasha. 2018. "The Mark of a Woman's Record: Gender and Academic Performance in Hiring." *American Sociological Review*, 83(2): 331–360.

Quillian, Lincoln, John J. Lee, and Brandon Honoré. 2020. "Racial Discrimination in the U.S. Housing and Mortgage Lending Markets: A Quantitative Review of Trends, 1976–2016." *Race and Social Problems*, 12(1): 13–28.

Radford, Alec, and Karthik Narasimhan. 2018. "Improving Language Understanding by Generative Pre-Training."

Radford, Alec, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners."

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2019. "Mitigating bias in algorithmic employment screening: Evaluating claims and practices." *arXiv preprint arXiv:1906.09208*.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating bias in algorithmic hiring." ACM.

Rambachan, Ashesh. 2022. "Identifying Prediction Mistakes in Observational Data." , (27111).

Rambachan, Ashesh, and Jonathan Roth. 2019. "Bias In, Bias Out? Evaluating the Folk Wisdom."

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig. 2020. "An Economic Approach to Regulating Algorithms." National Bureau of Economic Research Working Paper 27111.

Rao, Gautam. 2019. "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools." *American Economic Review*, 109(3): 774–809.

Raymond, Elora Lee, Ben Miller, Michaela McKinney, and Jonathan Braun. 2021. "Gentrifying Atlanta: Investor Purchases of Rental Housing, Evictions, and the Displacement of Black Residents." *Housing Policy Debate*, 31(3-5): 818–834.

Raymond, Elora Lee, Richard Duckworth, Benjamin Miller, Michael Lucas, and Shiraj Pokharel. 2018. "From Foreclosure to Eviction: Housing Insecurity in Corporate-Owned Single-Family Rentals." *Cityscape*, 20(3): 159–188.

Raymond, Elora Lee, Richard Duckworth, Benjmain Miller, Michael Lucas, and Shiraj Pokharel. 2016. "Corporate Landlords, Institutional Investors, and Displacement: Eviction Rates in Singlefamily Rentals." *FRB Atlanta community and economic development discussion paper*, , (2016-4).

Reagans, Ray, and Ezra W. Zuckerman. 2001. "Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams." *Organization Science*, 12(4): 502–517.

Redfin. 2023*a*. "Housing Investors Sell 1 in 7 Homes at a Loss—Highest Share Since 2016." *https://www.redfin.com/news/investor-homes-sold-at-a-loss/*.

Redfin. 2023*b*. "Investor Home Purchases Fell a Record 49% in the First Quarter." *https://www.redfin.com/news/investor-home-purchases-q1-2023/*.

Rejwan, Idan, and Yishay Mansour. 2019. "Top-k Combinatorial Bandits with Full-Bandit Feedback."

Rigollet, Philippe, and Assaf Zeevi. 2010. "Nonparametric Bandits with Covariates."

Rivera, Lauren A. 2012. "Hiring as Cultural Matching: The Case of Elite Professional Service Firms." *American Sociological Review*, 77(6): 999–1022.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data."

Journal of the American Statistical Association, 90(429): 106–121. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Roose, Kevin. 2023. "A Conversation With Bing's Chatbot Left Me Deeply Unsettled." *The New York Times*.

Rosen, Sherwin. 1981. "The Economics of Superstars." *The American Economic Review*, 71(5): 845–858.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision*, 115(3): 211–252.

Russo, Daniel, and Benjamin Van Roy. 2015. "An Information-Theoretic Analysis of Thompson Sampling."

Salzman, Diego, and Remco C.J. Zwinkels. 2017. "Behavioral Real Estate." *Journal of Real Estate Literature*, 25(1): 77–106.

Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. 2019. "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model."

Schumann, Candice, Zhi Lang, Jeffrey S. Foster, and John P. Dickerson. 2019. "Making the Cut: A Bandit-based Approach to Tiered Interviewing."

Shen, Ruoqi, Liyao Gao, and Yi-An Ma. 2022. "On Optimal Early Stopping: Over-informative versus Under-informative Parametrization."

Si, Nian, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. 2020. "Distributional Robust Batch Contextual Bandits."

Slivkins, Aleksandrs. 2014. "Contextual Bandits with Similarity Information." 36.

Smith, Julie M. 2021. "Algorithms and Bias." *Encyclopedia of Organizational Knowledge, Administration, and Technology*, 918–932.

Sterling, Adina D., and Roberto M. Fernandez. 2018. "Once in the Door: Gender, Tryouts, and the Initial Salaries of Managers." *Management Science*, 64(11): 5444–5460.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199.

Svenson, Ola. 1981. "Are we all less risky and more skillful than our fellow drivers?" *Acta Psychologica*, 47(2): 143–148.

Syverson, Chad. 2011. "What Determines Productivity?" *Journal of Economic Literature*, 49(2): 326–65.

Taniguchi, Hiroya, and Ken Yamada. 2022. "ICT Capital-Skill Complementarity and Wage Inequality: Evidence from OECD Countries." *Labour Economics*, 76: 102151. arXiv:1904.09857 [econ, q-fin].

Tessum, Christopher W., David A. Paolella, Sarah E. Chambliss, Joshua S. Apte, Jason D. Hill, and Julian D. Marshall. 2021. "PM_{2.5} polluters

disproportionately and systemically affect people of color in the United States." *Science Advances*, 7(18): eabf4491.

The Department of Housing and Urban Development. 2023. "Legislative History of Lead-Based Paint."

The White House. 2009. "Memorandum on Transparency and Open Government."

The White House. 2022. "The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America." The White House.

Todd, Sarah. 2019. "People are terrible judges of talent. Can algorithms do better?" *Quartz*.

Treisman, Anne M., and Garry Gelade. 1980. "A feature-integration theory of attention." *Cognitive Psychology*, 12(1): 97–136.

United States Court of Appeals for the First Circuit. 2020. "Students for Fair Admissions v. President and Fellows of Harvard College." cert. granted, 142 S. Ct. 895 (2022).

Upson, James, and Robert A. Van Ness. 2017. "Multiple Markets, Algorithmic Trading, and Market Liquidity." *Journal of Financial Markets*, 32: 49–68.

U.S. Census Bureau. 2021. "S2504 Physical Housing Characteristics for Occupied Housing Units."

US EPA, OAR. 2014. "Lead's Impact on Indoor Air Quality." *https://www.epa.gov/indoor-air-quality-iaq/leads-impact-indoor-air-quality*.

US EPA, OCSPP. 2013. "Lead Renovation, Repair and Painting Program." *https://www.epa.gov/lead/lead-renovation-repair-and-painting-program*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv:1706.03762 [cs].

Wang, Lu, Andrea Rotnitzky, and Xihong Lin. 2010. "Nonparametric Regression With Missing Outcomes Using Weighted Kernel Estimating Equations." *Journal of the American Statistical Association*, 105(491): 1135–1146. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/jasa.2010.tm08463.

Whatley, Warren C. 1990. "Getting a Foot in the Door: "Learning," State Dependence, and the Racial Integration of Firms." *The Journal of Economic History*, 50(1): 43?66.

Wheaton, David. 2023. "Fighting Appraisal Bias: How the Government and Housing Industry Can Better Address This Discriminatory Practice." *https://www.naacpldf.org/appraisal-algorithmic-bias/*.

Whittle, Richard, T. Davies, Matthew Gobey, and John Simister. 2014. "Behavioural Economics and House Prices: A Literature Review."

Yala, Adam, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. 2019. "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction." *Radiology*, 292(1): 60–66. PMID: 31063083.

Yu, Martin, and Nathan R. Kuncel. n.d.. "Pushing the Limits for Judgmental Consistency: Comparing Random Weighting Schemes with Expert Judgments." *Personnel Assessment and Decisions*, 6.

Zhang, Lei, and Tammy Leonard. 2021. "External Validity of Hedonic Price Estimates: Heterogeneity in the Price Discount Associated with Having Black and Hispanic Neighbors." *Journal of Regional Science*, 61(1): 62–85.

Zhao, Shuyin. 2023. "GitHub Copilot now has a better AI model and new capabilities."

Zillow. 2023. "What Is a Zestimate? Zillow's Zestimate Accuracy."

Zolas, Nikolas, Zachary Kroff, Erik Brynjolfsson, Kristina McElheran, David Beede, Catherine Buffington, Nathan Goldschlag, Lucia Foster, and Emin Dinlersoz. 2020. "Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey." Center for Economic Studies, U.S. Census Bureau Working Papers 20-40.