THE UNIVERSITY OF CHICAGO

ESSAYS ON IMPERFECT HUMANS AND IMPERFECT ALGORITHMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
DIAG DAVENPORT

CHICAGO, ILLINOIS
AUGUST 2022

To Winniebell East and Alberta Jett who paid for my crown;

David Davenport and Beverley East who taught me to wear it;

and Dupree Davenport who had the vision all along

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

One purported goal of research is to make sense of the world. Unfortunately, the past four years have been a particularly hard time to make sense of anything. So many lives have been taken for reasons I'll never understand. I've watched innocent lives be taken in police violence and gang violence in city streets across America, in and around classrooms, and in homes and hospitals on a nonstop cycle. Amidst this, I've been asked to learn what previous researchers have written about why people do the things they do. The hardest part of this process has been studying the world from the lens of an academic while experiencing the world as a person. And I fear often that I've developed my skills as a researcher at the expense of my ability to feel and be human. I am most thankful for the village that has carried me to this point and everyone who has kept my spirit afloat while navigating this tension. I hope one day I can do justice to my gratitude in person; perhaps these short words can suffice in the meantime.

Thank you Beverley East and David Davenport for setting this all into motion and giving me the superpower to see myself on my own terms and no one else's.

Thank you to Dupree Davenport, Chris Lorrain, Jamie Campbell, Adam Finn, and Kimberly Sellers who have been role models and intellectual lampposts through different phases of my journey here. The hope that one day I can inspire at least one person in the ways that you have impacted me has sustained me.

Thank you to all my friends who I've met on their own PhD paths, especially Sam Hirshman, Dan Medvedev, Alex Moore, David Munguia Gomez, Kariyushi Rao, Donovan Rowsey, Gülin Tuzcuoğlu, Kristina Wald, and Yuji Winet. Your collegiality and friendship and perspective have sustained me through these trying years.

Fern Ramoutar, I thank you for giving me faith that there can be rigor and unrelenting

humanity.

I thank all my friends for countless calls and messages and wishes, especially Assaph Aharoni, Jemar Bather, Kendall Brown, Michael Capehart, Carshena Culmer, Damian Gates, Bob and Jan Goldberg, Chantel Harley, Rubain Henry, and Antonio Sanders. The love and genuineness in our bonds have sustained me.

Portia Taylor, your love and partnership has sustained me.

Theresa Gregoire, your curiosities and insights have sustained me.

I thank all the folks at the Northwestern Prison Education Program, especially my students at Stateville Correctional Center whose letters and words and wisdom have sustained me. Your genuine thirst for knowledge has sustained me. The opportunity to serve and learn with you has sustained me.

I am forever indebted to Jens Ludwig and everyone at Crime Lab New York, especially Oludamilare Aboaba, Ellen Dunn, Sibella Matthews, Carolyn Silverman, and Greg Stoddard. The opportunity to learn from you all and play a small role in your work has sustained me.

Thank you to the faculty, especially Reid Hastie, Alex Imas, Jane Risen, Alex Todorov, and George Wu for helping shape my thinking and making the University and the Department a welcoming intellectual home for me.

To Betsy Levy Paluck, Sendhil Mullainathan, Devin Pope, and Richard Thaler: thank you for making this struggle worthwhile. Each of you has helped to materialize my hope of using big ideas to help shape the world. Both through your own work and through your feedback, I've honed in on a way forward that I can be proud of. It's a feeling I don't experience very often. You all have made me feel empowered and capable of bringing my boldest ideas to fruition.

Devin, thank you for fielding hundreds of ill-fated ideas and for holding me to a standard that I will forever be proud of. I will always appreciate the candor and carefulness in your critiques. And though my graphs may never look like yours, your standards will always be part of my internal dialogue.

Sendhil, thank you for having confidence in me and giving me confidence in myself. More than anything, your encouragement and wisdom has transformed a fuzzy dream to a real path. Your research, your advising, your impact, your vision, (though definitely not your schedule) all embody parts of the highest vision I have for myself.

I thank God and all the energy that has been ordained to pass me by for helping me see this through. I'll rest as much as I can. And then the real work will begin.

ABSTRACT

The most useful work in behavioral economics documents costly mistakes that people make in important settings and identifies ways to help them make better decisions. This dissertation presents four articles with those two goals in mind. In chapter one, co-authored with Yuji Winet, I examine perverse motivations people have in jury decisions and group decisions in general. In chapter two, I examine the efficiency of venture capital investing. In chapter three, co-authored with Amanda Agan, Jens Ludwig, and Sendhil Mullainathan, I examine how human biases interact with algorithmic biases in recommendation systems. In chapter four, I explore how police officers use algorithms for informative and performative purposes.

INTRODUCTION

In chapter one, co-authored with Yuji Winet, I study the psychology of group decisions. Many important social and policy decisions are made by small groups of people (e.g., juries, college admissions officers, corporate boards) with the hope that a collective process will yield better and fairer decisions. In many instances, it is possible for these groups to *fail* to reach a decision by not garnering a minimum number of votes (e.g., hung juries). Our research finds that pivotal voters vote to avoid such decision failure—voters who can "tip" their group into a punishment decision will be more likely to do so. This effect is distinct from well-known social pressures to simply conform with others or reach unanimity. Using observational data from Louisiana court cases, we find a sharp discontinuity in juries' voting decisions at the threshold between indecision and conviction (Study 1). In a third-party punishment paradigm, pivotal voters were more likely to vote to punish a target than non-pivotal voters, even when holding social information constant (Study 2), and adopted harsher views about the target's deservingness of punishment (Study 3). Using vignettes, we find that pivotal voters are judged to be differentially responsible for the outcomes of their votes—those who 'block' the group from reaching a punishment decision are deemed more responsible for the outcome than those who 'fall in line' (Study 4). These findings provide insight into how we might improve group decision-making environments to ensure that their outcomes accurately reflect group members' actual beliefs and not the influence of social pressures.

In chapter two, I examine the efficiency of venture capital investing. To study this question I combine a novel data set of over 16,000 startups (representing over $9 billion in investments) with machine learning methods to evaluate the decisions of early-stage investors. By comparing investor choices to an algorithm's predictions, I show that approximately half of the investments were predictably bad—based on information known at the time of investment, the predicted return of the investment was less than readily available outside options. The cost of these poor investments is 1,000 basis points, totaling over $900 million

in my data. I provide suggestive evidence that over-reliance on the founders' background is one mechanism underlying these choices. Together the results suggest that high stakes and firm sophistication are not sufficient for efficient use of information in capital allocation decisions.

In chapter three, co-authored with Amanda Agan, Jens Ludwig, and Sendhil Mullainathan, I explore how human biases interact with algorithmic biases. Conventional wisdom suggests algorithms trained on data about human behavior will mirror people's biased preferences. We argue here this *understates* the problem of algorithmic bias. Psychology shows our behaviors reflect not only our conscious preferences, but our implicit biases as well—particularly for quick, automatic decisions of the sort that are common online. Algorithms thus learn aspects of our worst selves that our conscious selves would repudiate. We show in both theory and experiment how this presumption leads algorithms astray: they wind up not just mirroring our biases, but exaggerating them. We focus on algorithms that help curate choice sets for users, where the learning of user's implicit bias creates a *double penalty* for out-group content: the algorithm down-ranks out-group content (so it is less likely to be seen), which compounds user bias against the out-group content they do see. The relevance of this finding is suggested by our audit of Facebook's algorithms across two countries, the US and India, which reveals a pattern of bias that is consistent with our model and that is quantitatively large. The underlying problem is widespread and requires rethinking how algorithms are built and deployed.

Finally, in chapter four, I explore how police officers use algorithms for informative and performative purposes. An enormous amount of work has studied the influence of information by exogenously providing it, but surprisingly little work has studied how and when people endogenously seek valuable information. To study the acquisition of information, I investigate an important setting where information is freely available—when New Jersey police officers book a defendant, they have complete discretion over whether to consult an algorithmic risk

score that predicts the defendant's likelihood of failing to appear in court as well as the defendant's likelihood of being rearrested if released. I find that officers frequently choose not to look at information that's free, simple, and easily available. Moreover, the selective viewing is far from random. Controlling for underlying risk, officers are more likely to consult the risk score for black defendants. Then, once the risk scores are seen, officers are more likely to issue warrants for black defendants, again controlling for risk. The pattern is consistent with officers strategically using the algorithm to signal race-neutral motives either to themselves or other actors in the criminal justice system. This behavior comes at a significant social cost. If officers simply followed the algorithm's recommendations, the warrant race gap would be reduced, NJ's jail population would be reduced, and the new crime rate would be reduced. I conclude by discussing policy implications for automation, decision aids, and human override in sensitive policy areas.

Any errors are my own.

CHAPTER 1

PIVOTAL VOTING: THE OPPORTUNITY TO TIP GROUP DECISIONS SKEWS JURIES AND OTHER VOTING OUTCOMES

1.1 Introduction

Important social and policy decisions are often determined by groups of people with the expectation that using collective decision processes will yield better and fairer decisions (Bang and Frith, 2017). Boards of Education decide what policies and practices public school systems adopt across the country; the Federal Open Market Committee decides how interest rates and the money supply in the US will be handled; admissions offices evaluate whether or not applicants will be admitted to their universities; and juries in court cases throughout the US deliver verdicts that determine whether defendants will serve time in prison or walk free.

An important feature of many group decision making processes is that, in order to arrive at a decision, a minimum number of voters must agree—a voting threshold must be crossed. One common example of a voting threshold is a simple majority (i.e., 50% + 1 vote), yet there are many other voting threshold rules as well. For example, some US congressional decisions require a two-thirds majority, and most criminal jury trials in the US require unanimity to deliver a verdict. In instances where the voting threshold is something other than a simple majority or when an even number of voters are evenly split among two factions (i.e., a tie), it is possible for groups to fail to decide (e.g., as happens with hung juries). The current research examines how this possibility of indecision may sway individual voters toward voting for outcomes that conflict with the conclusions they would naturally reach.

In 2020, the US Supreme Court took up the issue of setting jury voting thresholds for criminal trials. Prior to 2020, most states used unanimous voting thresholds, while some

used non-unanimous thresholds, as with Louisiana. The court ultimately decided to outlaw non-unanimous voting thresholds across the United States—and did so based in part on the seemingly innocuous but critical assumption that jurors vote for what they believe based on the evidence and deliberations, independent of where the voting threshold lies.[1] The present research calls this view into question. We find that how group members vote is dramatically influenced by the relative proximity of the group's current vote to the minimum number of votes needed to reach a final decision—to our knowledge, this topic has not been explored by previous research.

To understand this pattern of group behavior, imagine two juries that determine whether or not a defendant should be found guilty. Each jury consists of 12 jurors. Importantly, for each jury, there is a minimum threshold whereby some minimum number of jurors (e.g., 10) must agree in order to render a final verdict. If fewer than this number of jurors agree on a verdict, then the jury hangs and the fate of the defendant will depend on whether the prosecutor decides to retry the case with another jury[2] and on what verdict that next jury would deliver (Hannaford-Agor et al., 2002).

In the first jury, suppose the vote at a given moment is 9 votes to convict (and 2 to acquit), and the minimum voting threshold is 10 votes.[3] Alexis has yet to vote. As she is deciding her position, she holds the pivotal vote that determines whether the jury will render a verdict at all. How will she vote? Certainly, she may feel social pressure to join the majority by voting to convict, especially if she feels a desire to affiliate with that faction.

1. See Ramos v. Louisiana, 140 S. Ct. 1390, 206 L. Ed. 2d 583, 590 U.S. (2020)

2. In practice, the rate at which hung jury cases are retried is not well-estimated. However, Hannaford-Agor et al. (2002) indicate the rate at which mistrials (due to jury deadlock) are retried to a new jury is 32%. Cases that are not retried with a new jury either resolve as plea agreements (31.8%), are dismissed (21.6%), or are retried as a bench trial (2.4%).

3. Considering a snapshot is instructive even if the group does not vote sequentially in practice. For example, the snapshot could represent the forecast of votes a given voter has in mind while determining her own vote. Alternatively, in group decision settings with multiple rounds or where voters can see the votes of others and then change their minds, such a snapshot would speak to the moment between a given voter seeing the votes of others and determining her own vote.

Beyond this pressure to conform, however, she may also feel a desire for her group to merely reach a decision to avoid losing the opportunity to render a verdict at all. Both pressures steer her toward voting to convict. Now consider the second jury. In this jury, the current vote is also 9 votes to convict (and 2 to acquit), but here, the minimum voting threshold is 11 votes. Brianna has yet to vote. As she is deciding her position, she does not hold a pivotal vote and thus will not determine whether the jury will render a verdict at all. How will she vote? While she, too, may feel pressure to conform with the majority by voting to convict, she will be unlikely to vote for conviction simply to render a verdict because her vote cannot push the group vote across the minimum voting threshold. That is, regardless of whether she votes "guilty" or "not guilty," it is impossible for her to vote to pull the group out of a state of indecision (hung jury) and into a conclusive state of decision (a final verdict).

In this research, we will contrast the experiences of Alexis and Brianna, and explore the psychology of being a pivotal voter.[4] More specifically, we hypothesize that the opportunity for group-level decisiveness dramatically influences how pivotal voters vote, which skews group voting outcomes as a result.

In this paper, we present results from one observational dataset and three preregistered lab experiments. These studies provide evidence that pivotal voters are more likely to vote in favor of punishment than they otherwise would as non-pivotal voters, and that they appear to adopt harsher views in order to rationalize their votes. In Study 1, we use data from Louisiana court cases (N = 1,960) to show that juries that are subject to non-unanimous voting threshold rules disproportionately reach conviction verdicts by votes that just barely

4. A third scenario that one could consider is that of Christine. For her, the minimum voting threshold is 10 votes and the current vote is 10 votes to convict (and 1 to acquit). Again, she may feel pressure to conform and join the majority by voting to convict; however, her vote will have no impact on the final outcome because the jury will convict regardless of her vote (i.e., the threshold of 10 votes has already been reached). We find that the desire to produce a firm answer is greatly motivating—therefore, the fundamental motivations of Alexis and Brianna are different from those of Christine. We leave the psychology of holding out after the group is guaranteed to produce a firm answer to future research. The current research will sharply focus on comparing the state of indecision without the power to change it (Alexis) to the state of indecision with the power to change it (Brianna).

cross voting thresholds—a pattern consistent with the proposed effect and inconsistent with classic conformity pressure explanations. In Study 2a, we conceptually replicate Study 1 by conducting a lab experiment in which participants decide whether to punish another participant for selfish behavior in a third-party punishment paradigm, in which we find that pivotal voters were more likely to vote to punish than non-pivotal voters. In Study 2b, we use the same paradigm to test the likelihood of voting for acquittals and here we do not find that pivotal voters behave any differently than non-pivotal voters. In Study 3, we test a potential psychological mechanism underlying the effect by examining private beliefs. Not only are pivotal voters more likely to punish the target than non-pivotal voters but we find that they also convince themselves that the targets deserve it more. In Study 4, we further probe mechanism by examining how people evaluate pivotal voters, which may offer insight into pivotal voters' own subjective experiences. Participants read vignettes involving pivotal voters and make responsibility judgments about them. We find that pivotal voters are deemed to be more responsible for the outcome when their vote blocks the group from reaching a final verdict to convict (i.e., thus leading to a hung jury) versus when they "tip" the group into reaching such a verdict. Together, these findings suggest that pivotal voters may be more likely to both vote for punishment outcomes and update their beliefs correspondingly, not because of how they privately interpret the evidence or circumstances, but because they find it aversive to bear the responsibility of causing a "non-outcome" for their group.

Past research in several literatures make diverging predictions about how voters may behave in group decision making contexts where choices can lead to inconclusive outcomes. Social forces like conformity and polarization suggest that voters will seek to align their votes with others, leading groups to either vote unanimously for a given outcome (Asch (1955), Cialdini and Goldstein (2004), Levy (1960)) or split into separate factions that unite around separate outcomes (Penrod, Pennington, and Hastie (1983), Isenberg (1986),

Lord, Ross, and Lepper (1979)). Aside from social pressures, the mere potential for group indecision can also have material impacts on group decisiveness—one possibility is that the option not to decide, or to defer, could push individual group members to prevent group decisions altogether. When making especially difficult decisions (e.g., judging a defendant's guilt), individuals can experience negative feelings associated with the stress of potentially making a regrettable choice (Dhar (1997), Luce (1998), Tversky and Shafir (1992)). In group decision making contexts, individual voters could cope with such stress by voting, when able, to maintain their group's state of indecision—that is, a pivotal voter could prefer a hung jury outcome that defers a decision to another jury, rather than allow themselves to be complicit in delivering what could be an unjust verdict.

Whereas social pressure accounts predict voting distributions that collect around the point of unanimity and indecision accounts predict distributions that collect somewhere below the minimum voting threshold, a motivation to meet a group's goals (such as by reaching a final decision) (Allen et al. (2017), Anik and Norton (2020), Rees-Jones (2018)) and avoid disagreement (Tuncel et al., 2016) would instead predict voting distributions that collect at the exact point where the minimum voting threshold is crossed—precisely where the minimal conditions for a conclusive outcome are met. Our research thus integrates three literatures which have conflicting predictions for group decisions where there is a possibility of decision failure: classic conformity (unanimity), deferral and avoidance of choice (decision aversion), and goal attainment (indecision aversion).

First, we acknowledge the longstanding finding that people are more likely to take actions taken by others (Asch (1955), Cialdini and Goldstein (2004), Levy (1960))—the well-documented effect of classic conformity pressure. Importantly, our proposed effect goes above and beyond this pressure. Controlling for conformity pressure, we provide evidence that in group decision making contexts the tension between preferring to avoid or defer a decision and preferring conclusive outcomes strongly favors conclusive outcomes

in punishment contexts. Further, we present evidence that suggests this behavior operates through a change in private attitudes resulting from a sense of obligation or expectations from the group, thus disentangling the normative and informational influences described by Deutsch and Gerard (1955). That is, despite voters' initial private desire for any given outcome, they are motivated to rationalize pulling groups out of a state of indecision and into a state of decision. This individual-level drive causes group voting distributions to cluster not around unanimity, nor around some point below the minimum threshold, but at the exact point at which a voting threshold for a decisive punishment is crossed.

In addition to these social mechanisms, we find some evidence of a cognitive mechanism underlying the proposed effect. The rich literature on the need for cognitive closure demonstrates that people often have a desire for knowledge or a firm answer (Kruglanski and Webster, 2018). Previous research has shown that this desire to close epistemic gaps can be conceptualized both as a stable personality trait as well as a motivational tendency that can be induced by a given decision making context (e.g., under time pressure; Webster and Kruglanski (1994), Kruglanski and Freund (1983), Heaton and Kruglanski (1991), Jamieson and Zanna (1989)). In particular, past work shows that need for closure has enormous consequences for how group members interact with each other, all in an effort to receive or acquire firm knowledge from the environment (e.g., by preferring autocratic leaders and silencing dissent; Kruglanski et al. 2006). Our research extends this literature by suggesting that group affiliation may heighten a need for closure which manifests as a need to produce a firm, decisive answer (e.g., to punish others). Taken together, we demonstrate that the effect of being a pivotal voter holds in high-stakes field settings and the laboratory, for a variety of moral decisions in a punishment context. In other work we also show that the effect holds true for objective decisions with financial incentives and is moderated by the degree of affiliation among group members. This work thus extends and generalizes findings from recent work on impasse aversion (Tuncel et al. 2016) beyond mere framing effects and negotiation contexts, and

provides further insight into the psychological mechanisms explaining why inherent value is placed on decision and agreement over indecision and impasse. For more details on each study, see Materials and Methods.

Our work has major implications for jury trials and group decision making contexts at large that use voting thresholds with the potential for indecision (e.g., non-simple-majority voting thresholds). In jury contexts, these findings suggest that many defendants who have been convicted (even by unanimous juries) may not have in fact been considered guilty by all members of the jury, for reasons beyond mere conformity. As a result, innocent people are likely being imprisoned due to the predictable and systematic influence of our proposed effect. These findings also extend beyond the criminal justice system—the possibility of indecision is fundamentally aversive for groups and their members in contexts of all kinds, especially when group members feel more connected to each other. These potential negative ramifications stand in opposition to the overwhelmingly positive literature on the benefits of affiliation amongst group members (Carmeli et al. 2009, Carmeli, Brueller, and Dutton 2009, Cameron and Dutton 2003, Fitzsimons, Sackett, and Finkel 2016). In these situations, much care must be taken to structure group dynamics (e.g., by blinding votes and increasing commitment to votes when submitted) to ensure that they trend towards voting outcomes that reflect group members' true beliefs, rather than trend toward conclusive outcomes for their own sake.

1.2 Study 1 Results

Our first study (N = 1,960), uses a large dataset of criminal jury trial outcomes in Louisiana where a unique feature of their legal system (prior to 2018) allows us to tease apart the effects of thresholds from conformity pressures: unanimity among jurors was not required

to reach group decisions (i.e., verdicts).[5] Instead, a minimum of 10 out of 12 jurors were required to agree in order to render a conviction or acquittal; failing to reach this minimum threshold resulted in a hung jury, which meant no decision was reached and the defendant might or might not go on to be retried in another trial. When the threshold is unanimity, it is impossible to distinguish between indecision aversion and conformity pressure, which makes Louisiana prior to 2018 an ideal setting to study our proposed effect. Consistent with the effect, and in sharp contrast to a classic conformity model, we find a significant discontinuity in the jury voting distribution at the exact point at which a conviction threshold is crossed (i.e., at 10 votes to convict). Indeed, less than 1% of Louisiana juries in our data conclude with 9 votes to convict (i.e., one fewer than the minimum number of votes to convict, which represents a hung jury) while 17% conclude with 10 votes to convict (i.e., the exact minimum to convict, which avoids a hung jury outcome; $\chi^2 = 91.18$, df $= 1$, p ¡ .001). This pattern suggests that pivotal voters disproportionately vote to tip group votes into conclusive decisions to convict. Moreover, remaining holdouts (non-pivotal minorities) often still resist unanimity pressure, resulting in decisive, but non-unanimous outcomes. These results provide evidence against a purely classic conformity account which would predict the most votes to amass at 12–0 votes to convict and no votes at 10–2 or 11–1, and an indecision attraction account which would predict votes to amass between 3–9 and 9–3 (inclusive) where no verdicts are possible. On the other hand, it isn't clear that such a discontinuity exists for acquittals: 2% of cases end with just enough votes for an acquittal while exactly 0% end just shy of the threshold.[6] These results also suggest an asymmetry in the pivotal voter effect–the motivation to tip only has an effect when on the cusp of convicting, not acquitting. Given the non-experimental nature of these data, however, these findings should be interpreted

5. In 2018, after the period we study, the people of Louisiana voted to require the unanimous agreement of jurors (12 out of 12 jurors), rather than the previous 10 of 12 jurors, to convict defendants charged with felonies.

6. We believe conducting formal inference is not meaningful given the small sample size (40 cases) and presence of a zero-cell, but descriptively there appears to be little evidence of a discontinuity.

with caution. We cannot rule out that the observed pattern is driven by confounds that are orthogonal to our effect, such as jury composition and the strength of evidence. To resolve these potential confounds, we further investigate this pivotal voting phenomenon in three experimental studies.

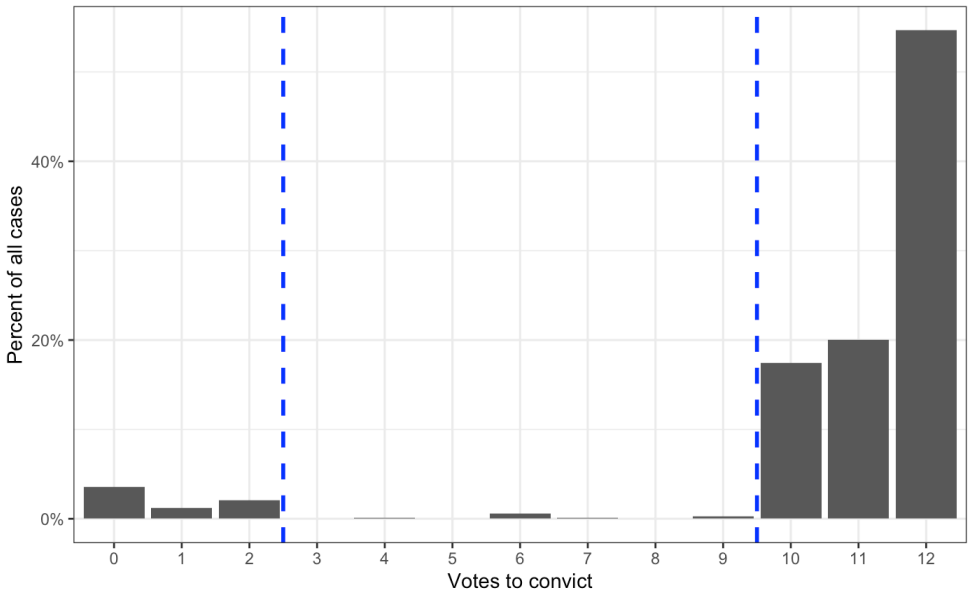

Figure 1.1. This figure depicts the percentage of total Louisiana juries in criminal court cases from 2011 to 2016 with a given final distribution of votes to convict (versus acquit). All juries consisted of 12 jurors. The blue dashed lines indicate the minimum voting threshold for a given outcome based on Louisiana's deliberation laws (during this time period)—two or fewer votes to convict resulted in a not guilty verdict; three to nine votes to convict resulted in a hung jury; ten or more votes to convict resulted in a guilty verdict.

1.3 Study 2a Results

In a second study (N = 261), we designed a test of the effect in a more controlled setting. We conducted a preregistered online experiment in which participants served as judges in a third-party punishment game. Participants were informed that they would be a part of a group with three other participants and would decide whether or not to punish another

participant playing the role of the dictator in a dictator game. To render a decision, the group had to reach a minimum threshold of either three votes (simple majority) or four votes (unanimity). Failure to reach the threshold meant the decision would ostensibly be handed off to another group. First, all participants learned that the dictator made a selfish decision (to share \$0.20 of a \$1.00 endowment and keep \$0.80 for themselves). Then, participants entered group deliberations, where they learned that the other group members voted 2–1 in favor of punishing the dictator. In the 4-vote-threshold condition, no group decision was possible regardless of how participants voted—thus making them a non-pivotal voter. In the 3-vote-threshold condition, participants were able to cast the decisive vote (for group punishment versus submitting "no answer")—thus making them the pivotal voter. A key benefit of this paradigm is our ability to vary whether participants were pivotal while holding the number of other group members' votes constant. This allows us to estimate the marginal impact of being a pivotal voter above and beyond classic conformity effects.

Our main dependent variable is the rate at which participants voted to punish the dictator (punish = 1, do not punish = 0). We find that being a pivotal voter led to a large difference in the probability of voting for punishment: 60.00% of pivotal voters punish while only 40.46% of non-pivotal voters punish ($\chi^2 = 9.20$, df = 1, $p = .002$). These results corroborate the findings from Study 1: the opportunity to cast a decisive vote increases one's willingness to vote with the majority faction and dole out punishment.

1.4 Study 2b Results

In a nearly-identical study as Study 2a, we designed a separate and complementary preregistered online experiment (N = 298) that investigated the effect of being pivotal for acquittal, as opposed to punishment. Participants were again assigned to 3- and 4-vote-threshold conditions, but unlike Study 2a, they were faced with a 2–1 vote in favor of acquitting the dictator. This meant participants were given an opportunity to either tip their group vote

into an acquittal outcome or to render no answer, causing the decision to be ostensibly handed off to another group. Contrary to Study 2a and our preregistered predictions, we find that pivotal voters are no more likely to vote to acquit: 74.21% of pivotal voters acquit while 73.24% of non-pivotal voters acquit ($\chi^2 = 0.03$, df = 1, $p = .870$).[7] Together, the results of Studies 2a and 2b demonstrate an asymmetry whereby the effect of being a pivotal voter can only stand to harm targets, which threatens the integrity of institutions that rely on group decisions to dole out fair punishments.[8]

1.5 Study 3 Results

In a preregistered third study (N = 1,633), we tested our hypothesis using group sizes that more closely resemble actual juries and we measured judgments about targets to assess whether belief-updating may explain why pivotal voters become more likely to vote for conclusive group decisions. We used a similar design as Study 2, in which we manipulated whether participants were pivotal or non-pivotal voters, except that we also increased the group size to 12 judges, we included three voting threshold conditions, and we measured beliefs about targets' deservingness of punishment. Participants were assigned one of three voting thresholds: their group would require at least either 10, 11, or 12 votes for a given option in order to deliver a verdict. As in Study 2, failure to reach a verdict would result in the final decision being deferred to another group. As in Study 2, participants were assigned to be either a pivotal voter or a non-pivotal voter. For pivotal voters, the group was always one vote away from crossing the voting threshold (e.g., in the 11-vote threshold condition, the group vote was 10–1 in favor of punishment), which meant voting to punish would

7. For the remainder of the paper, we focus on punishment outcomes specifically, given the disproportionate prevalence with which jury verdicts end in convictions, as opposed to acquittals. For example, 92.10% of trials in the Study 1 data end in conviction compared to the 6.79% that end in acquittal.

8. While punishment decisions are intrinsically important, we do not believe that the effect is unique to punishments. In supplemental studies omitted from this draft, we show the effect in a non-moral domain where we find a symmetric effect of tipping between two counterbalanced options trivia responses in an incentive-compatible, non-moral group trivia game.

lead to a punishment verdict and voting not to punish would lead to a decision deferral. For non-pivotal voters, the group was always two votes away from crossing the threshold (e.g., in the 11-vote threshold condition, the group vote was 9–2 in favor of punishment), which meant the group would defer the decision no matter how the non-pivotal voter voted. Importantly, we measured the degree to which participants believed the target deserved to be punished after learning how the target allocated the endowment, and we randomized whether participants reported this belief before or after learning how the other group members voted.[9] All participants report their beliefs before casting their own vote, which prevents their reported beliefs from being influenced by post-decision justification processes. We again find that participants are more likely to vote to punish a target when they are a pivotal voter, and that this effect persists across various voting thresholds. Interestingly, we also find that pivotal voters update their beliefs about targets' deservingness—this suggests that the social pressure pivotal voters feel to punish can drive them to deviate from the private attitudes they hold based strictly on the evidence.

9. Formally, the process by which real jurors initially learn the votes of others is through an initial ballot held at the beginning of deliberations, which may or may not be anonymous. However, researchers have shown that informal "predeliberations" occur where jurors share their own beliefs, learn the beliefs of others, and potentially update their own beliefs (Sandys and Dillehay 1995). Our experimental design departs from these realities and doesn't map perfectly onto the potentially public and dynamic nature of jury deliberations, but allows us to unpack the psychology that would be present throughout the process.

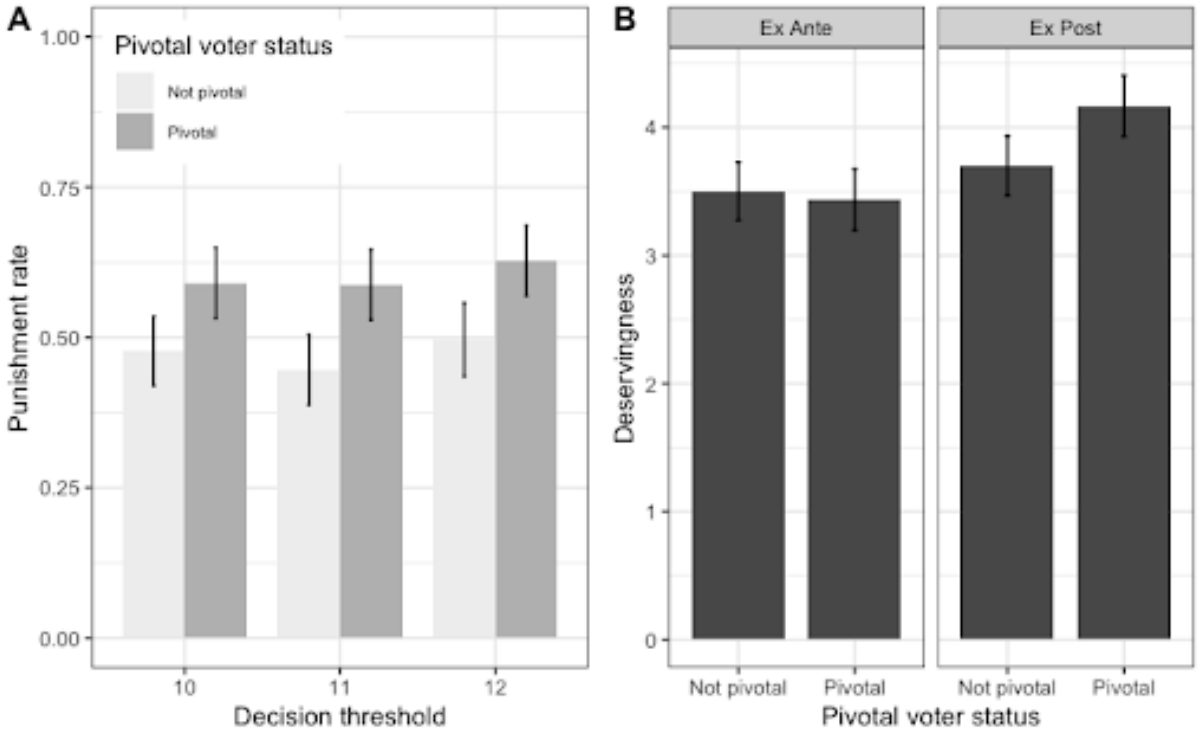

Figure 1.2. This figure depicts the rate at which participants chose to punish a selfish dictator in a Dictator Game (Panel A) and the extent to which participants believed the dictator deserved to be punished (Panel B). Participants were assigned to one of three "Decision threshold" conditions and were assigned to be pivotal voters (one vote away from the threshold) or non-pivotal voters (two votes away from the threshold). "Ex ante" and "Ex post" are relative to learning the social information (i.e., other votes). All beliefs are elicited after the details of 'the case' are known. Error bars depict 95% confidence intervals.

Figure 1.2a shows that the rate at which participants punish the dictator differs between situations in which they are pivotal (60%) and those in which they are non-pivotal (47%). However, to conduct formal hypothesis testing, it is important to unpack the differences between pivotal and non-pivotal voters while controlling for the effect of classic conformity. Accordingly, we first conduct a logistic regression of the participant's choice (voted to punish = 1; voted not to punish = 0) against a dummy variable for whether the participant is a pivotal voter (pivotal = 1; not pivotal = 0) and a variable containing the number of group votes assigned to punish the dictator. The parameter of interest is the coefficient on the pivotal voter dummy variable. We replicate our basic effect, finding that pivotal voters are

more likely to punish the dictator than non-pivotal voters ($\beta_{pivotal}$ = 0.46, SE = 0.12, p ¡ .001).

Figure 1.2b shows participants' beliefs about targets' deservingness of punishment before and after seeing other group members' votes. In the left panel of 1.2b we examine the difference in beliefs among participants who provided deservingness ratings before seeing the other members' votes (but after learning about 'the case'; ex-ante private beliefs).[10] The remaining participants provided their deservingness rating after seeing everyone else's vote (and after learning about 'the case'; ex-post private beliefs). These ex-post beliefs capture the potential influence of other votes. By design, ex-ante beliefs were equivalent between (soon-to-be) pivotal voters (M = 3.43, SD = 2.43) and non-pivotal voters (M = 3.50, SD = 2.37), t(813) = 0.40, p = .693, d = –0.03, and capture the unswayed, private beliefs that participants have about the target. By contrast, the ex-post beliefs were not equivalent between pivotal voters (M = 4.17, SD = 2.47) and non-pivotal voters (M = 3.70, SD = 2.40)—being pivotal causes participants to rate targets as more deserving of punishment, t(816) = –2.73, p = .006, d = 0.19. Moreover, the difference persists even when controlling for current votes ($\beta_{pivotal}$= 0.49, SE = 0.20, p = .016) (See Table 2). This result demonstrates that pivotal voters adopt harsher views towards the target. Importantly, the target's behavior is unambiguous (i.e., all participants are given the entire set of facts: that the dictator was endowed with $1.00 and only shared $0.20) and the number of current votes is controlled for, so this difference in beliefs cannot be readily explained by access to different factual or social information. Instead, we interpret this result as evidence that pivotal voters' attitudes incorporate the influence of others—they adopt the ex-post belief that is congruent with their increased propensity to punish, but incongruent with their independent conclusions about the facts and the (socially influenced) beliefs they would have had if not pivotal.

10. Here we label ex-ante and ex-post relative to learning the social information (i.e., other votes). All beliefs are elicited after the details of 'the case' are known.

1.6 Study 4 Results

In our final study (N = 505), we explore attributions of pivotal voters' causality, accountability, and responsibility for group outcomes as a function of whether or not their vote blocks a group decision (e.g., leads to a hung jury) or enables one (e.g., leads to a verdict). All participants read 10 vignettes that each describe a scenario in which a victim was harmed, and another party carries some ambiguous amount of responsibility for causing that harm. For each vignette, participants learn that a jury with 12 members is deliberating on whether to punish the other party ("defendant"), and that there is a final voter who must decide whether to vote to punish or acquit. Participants are assigned to one of two voting threshold conditions: either they learn that the juries have a majority-rules threshold, meaning 7 out of 12 votes are needed to convict, and that each jury's current vote is 6–5 to convict, or a unanimous threshold, meaning 12 out of 12 votes are needed to convict and that the current vote is 11–0 to convict. In all situations, the final voter is pivotal. For each vignette, for each subject, we randomize the voting decision outcome: the pivotal voter either votes against the majority to acquit—thus leading to a hung jury—or with the majority to convict—thus leading to a final verdict to convict. We then ask the participant 3 questions (Likert scale, 1-7) capturing the degree of responsibility, accountability, and causation they attribute to the pivotal voter with respect to the group's ultimate outcome. This study allows us to explore the likely subjective experience driving the normative influence and speaks to why voting against the group as a pivotal voter may be so aversive.

We average the three responses together to generate an accountability index (alpha = .96). We conduct an OLS regression analysis with the accountability index as the dependent variable, and the voting decision outcome (hung jury = 1, punishment verdict = 0) and voting threshold as independent variables (unanimous threshold = 1, majority threshold = 0). We find that with a majority-rules voting threshold, pivotal voters are considered slightly more responsible for the group outcome when they vote against the majority ($\beta_{\text{hung jury}} = 0.29$,

SE = 0.12, p = .021). However, when under a unanimous threshold, the effect is dramatically larger ($\beta_{\text{hung jury x unanimous}} = 1.53$, SE = 0.17, $p < .001$). These results, captured in Figure 1.3, demonstrate how people view pivotal voters who block or enable group decisions and suggest how pivotal voters themselves may anticipate or internalize additional responsibility. This possibility suggests at least one pathway that leads pivotal voters to avoid indecision.

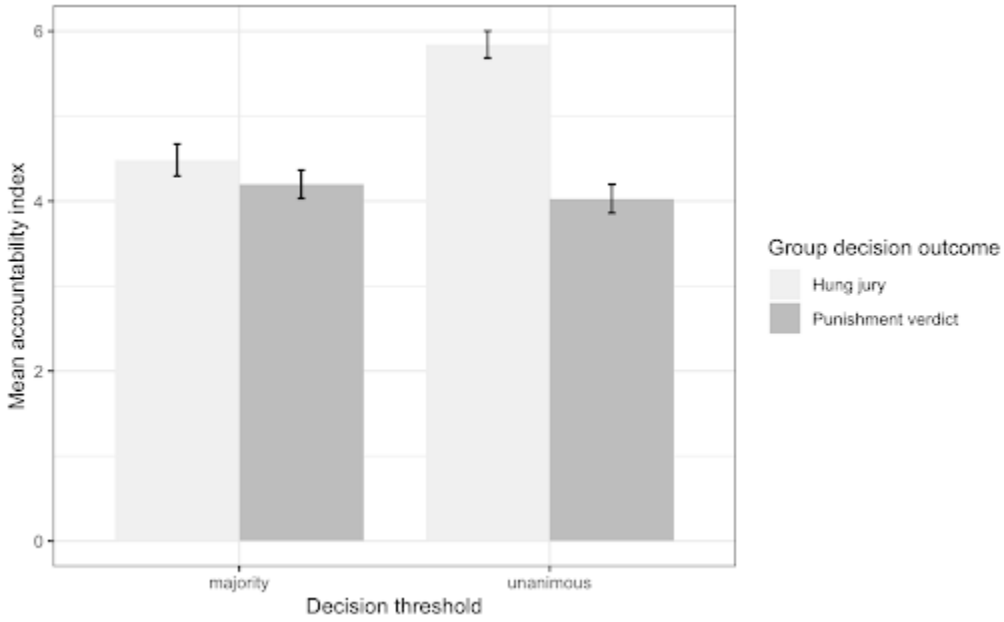

Figure 1.3. This figure depicts participants' responsibility judgments of pivotal voters whose choices led to a hung jury or a punishment verdict. The pivotal voter was either a member of a majority-rules jury or a unanimous-threshold jury. Error bars depict 95% confidence intervals.

1.7 Discussion

Groups are responsible for making many of society's most important decisions; despite this reality, past research has shown that group decision making can suffer from many process failures. We find evidence for yet another process failure for groups with complex decision rules: the mere opportunity to make conclusive group decisions dramatically influences pivotal voters and ultimately skews group voting outcomes. Our findings provide new insights into a novel group decision making process failure—rather than accept their group's failure to reach a decision, pivotal group members opt to simply to reach a conclusive outcome.

This is especially concerning because one could argue that votes ought to align with the beliefs the voters would hold if not pivotal.

Jury trials help to illustrate the adverse real-world consequences of these findings. In a recent Supreme Court case, it was argued that the 735 (37%) defendants convicted by non-unanimous juries in Louisiana between 2011 and 2016 would have received hung juries had the threshold merely been set to unanimity. Our studies, by contrast, suggest that moving the threshold would have been less effective than the Supreme Court suggests because jurors would continue to be motivated to reach a conclusive group decision to convict (versus becoming a hung jury), leading the group voting distribution to adjust to meet the minimum voting threshold for conviction regardless of where that threshold is set. We posit that changing the jury threshold, which occurred as a result of this Supreme Court case, was neither necessary nor sufficient—rather, other remedies must be called for to reduce unconstitutional convictions.

Our research suggests that groups of all kinds may be subject to a fundamental psychological aversion that causes them to prefer mere decisiveness over adhering to their own beliefs about the matter at hand. In the real world there are institutional and personal influences that compound or exacerbate the tendency to vote against one's true beliefs. For example, courts frequently levy "dynamite charges," whereby judges regularly encourage deadlocked jurors to work harder to reach a decision (Smith and Kassin 1993). In addition, more basic private preferences to "just be done with it" may lead to false consensus if voters are not adequately motivated to deliberate. In our lab studies, we isolate our proposed effect from institutional influences (by omitting them) and personal influences (via randomization), but expect that these elements merely amplify our effect in the real world.

With this in mind, we underscore the need for care in designing choice architectures and avoiding "sludge" (Thaler 2018) that exacerbates psychological biases such as conformity pressures and the effect of being a pivotal voter. An open question is what interventions

might effectively mute this effect across the many settings in which groups make impactful decisions. We believe that the strong moderating effect of the asymmetry in responsibility judgments along with the interpersonal affiliation seem to provide fruitful directions.

Our findings raise a number of other interesting questions for future research. Firstly, does this finding extend to larger groups, such as the U.S. Congress and the United Nations? Groups like these are often tasked with even more consequential decisions, so the heightened stakes may make indecision more aversive. However, larger groups may also be subject to other pressures and types of expertise that crowd out the effect (e.g., party loyalty in political decisions). Secondly, future research should determine whether the assigned location of the threshold itself may signal different degrees of the decision's importance to group members. For example, voters may infer that a decision that requires a simple majority is less consequential than one that requires unanimity. Thirdly, what explains the reason we find this effect in contexts as varied as punishment decisions and non-moral, incentive-compatible trivia games, but not acquittals? One potential answer is that punishing and submitting trivia answers are acts of commission, whereas acquitting—neglecting to punish someone—is an act of omission, and previous research has documented asymmetries between the two types of acts (Spranca, Minsk, and Baron 1991). Lastly, we suspect that individual differences vary the magnitude of the effect. We encourage future research to explore the extent to which gender, culture, personality traits, and group composition moderate the findings.

The list of barriers to effective collaboration is ever-expanding. Although there are many benefits to group decision making, the drawbacks should not be ignored. Our research demonstrates the influence of voting thresholds on pivotal voters and group outcomes, and provides a warning signal for how to encourage group members to vote for what they truly believe: design choice environments that commit members to their beliefs more than to crossing thresholds.

1.8 Materials and Methods

This research was approved by the University of Chicago Institutional Review Board (IRB 19-1060). All participants provided informed consent.

1.8.1 Study 1

In order to investigate the ramifications of Louisiana's non-unanimous voting rules, *The Advocate*, Louisiana's largest daily newspaper, collected data from 75% of all jury trials in the state between 2011 and 2016. This extensive effort was part of a thorough journalistic investigation into the causes and consequences of this unique system. The publication has made this data set publicly available.[11]

The Advocate notes, "...Of the cases in the data set, it was possible to determine whether verdicts were unanimous on 993 convictions out of the 2,027 cases that ended with at least one guilty verdict from a jury. Those cases cover half of the state's 64 parishes, though they are heavily weighted toward the large parishes of Orleans, Jefferson, St. Tammany, East Baton Rouge and Caddo. Collectively, those parishes are responsible for about 68 percent of the convictions in the state and roughly 69 percent of the data on jury unanimity in the data set."

In our analysis, we first limit the sample to those 3,794 charges with 12-person juries (70.3% of the dataset), to exclude charges with unconventional jury sizes. We then further subset the data to only include jury outcomes for which The Advocate was able to collect information about individual juror's votes. Out of 3,794 outcomes, 1,960 (51.7%) had data on the votes of each juror, yielding 1,960 jury outcomes across 1,044 trials. A wide variety of charges are represented in the resulting data, but the three most common charges demonstrate the gravity of the charges considered: 2nd degree murder (20%), armed robbery

11. See Advocate Staff Report (2018, April 1). *Tilting the scales series: Everything to know about Louisiana's controversial 10-2 jury law.* Nola. https://www.nola.com/news/courts/article_64f67fc8-9ab4-56b6-bb45-598b6795cffa.html

(8%), and firearm possession (7%). Based on discussions with staff from *The Advocate*, we do not believe the jury outcomes are correlated with the probability of appearing in the sample, which would introduce significant sampling bias. However, we are not able to directly observe or test this belief.

1.8.2 Study 2a

This study used a 2-condition (voting threshold: majority rule vs. unanimity rule) between-subjects design to test whether participants would be more likely to vote to punish a third-party if they held the decisive vote.

Participants and Procedure

We initially set out to recruit 300 participants from a community population for a preregistered virtual lab study in return for \$1.40 (M_{age} = 27.6; SD_{age} = 11.24; 70% female; all data and materials are available on OSF[12]). Due to COVID-related complications, we were only able to recruit 261. Participants learned they would be one of four judges who would decide, as a group, whether or not to punish another participant playing the role of the Dictator in an economic Dictator Game (Camerer 2011) with yet another participant. The Dictator had been granted an endowment of \$1.00 and made a selfish decision to allocate \$0.80 to themselves.

Participants were randomly assigned to one of two voting threshold treatments: either three votes (majority) or four votes (unanimity) were necessary to reach a group decision. Participants learned that a failure to reach a group decision meant the final decision would be left to a different group of judges. Participants read that a group decision to punish resulted in a deduction of \$0.19 from the Dictator's bonus and that a decision to not punish would allow the Dictator to keep the full \$0.80. Participants completed three comprehension checks

12. https://tinyurl.com/28vwyuc8

and one attention check to ensure they understood the rules of voting and punishment, and that they were paying attention.

Participants next observed the ostensible results of the Dictator game. They learned that two other judges had voted to punish and one judge had voted not to punish. Participants then made their ruling: "Do not punish" versus "Punish by subtracting 19 cents." As they made this decision, participants could see how the other group members voted and what the final group decision would be depending on how they cast their own vote—in the 3-vote-threshold condition, voting to punish resulted in punishment while voting not to punish resulted in decision deferral; in the 4-vote-threshold condition, voting always led to deferral. The answer choices, presentation of group votes, and indication of what final group decisions would be reached if the participant chose a given answer were yoked and presented in counterbalanced order. After making their decisions, participants reported on their subjective experiences of the ease of making their decision, the perceived influence of the other group members, and their own satisfaction with the final group decision. Participants also reported the perceived gender of the dictator. These results are discussed in omitted supplemental work.

1.8.3 Study 2b

This study used a 2-condition (voting threshold: majority rule vs. unanimity rule) between-subjects design to test whether participants would be more likely to vote to acquit a third-party if they held the decisive vote.

Participants and Procedure

We requested 300 participants through Prolific Academic, for a preregistered online study in return for \$1.15. This process returned 298 participants ($M_{age} = 34.01$; $SD_{age} = 12.69$; 60.74% female).

This study was designed to be identical to Study 2a, except that participants learned that two of the other judges had voted not to punish and one judge had voted to punish (unlike Study 2a, where one judge voted not to punish and two judges voted to punish). As in Study 2a, participants could see how the other group members voted and what the final group decision would be depending on how they cast their own vote. However, given this voting distribution, participants in the three-vote-threshold condition saw that voting not to punish resulted in an acquittal outcome while voting to punish resulted in decision deferral, and participants in the four-vote-threshold condition saw that voting always led to deferral. Results of subjective experience measures and perceived gender of the dictator are discussed in omitted supplemental work.

1.8.4 Study 3

This study used a 2 (pivotal voter status: pivotal vs. not pivotal, between subjects) x 3 (voting threshold: 10 vs. 11 vs. 12 votes, between subjects) x 2 (belief elicitation: before vs. after, between subjects) design to both test whether pivotal voters update their private beliefs as they become more likely to vote for decisive outcomes, and whether the effect persists across different voting thresholds.

Participants and Procedure

We requested 1,600 "Cloud Approved" participants through Cloud Research, for a preregistered online study in return for $1.00. This process returned 1,633 participants (M_{age} = 41.00; SD_{age} = 13.03; 55.36% female).

Participants completed procedures that were similar to Study 2; however, we also increased the size of the group to 12 judges, we included additional voting threshold conditions, and we measured participants' beliefs about the target's deservingness of punishment. Participants were randomly assigned to one of three voting threshold conditions: at least 10 votes, 11

votes, or 12 votes were required for the group to deliver a verdict. As in Study 2, participants learned that failing to reach a verdict meant the decision would be left to another group. Importantly, participants were assigned to one of two pivotality conditions: either they were a pivotal voter or not a pivotal voter. For pivotal voters, their group's current vote (before casting their own individual vote) was always one vote away from crossing their minimum voting threshold for a punishment verdict (i.e., in the 10-vote threshold condition, the group vote was 9–2 in favor of punishment, in the 11-vote condition it was 10–1, and in the 12-vote condition it was 11-0). For non-pivotal voters, their group's current vote was always two votes away from crossing the threshold for a punishment verdict (i.e., in the 10-vote threshold condition, the group vote was 8–3 in favor of punishment, in the 11-vote condition it was 9–2, and in the 12-vote condition it was 10-1). By varying both the number of current votes and the minimum voting threshold, we are able to directly compare pivotal and non-pivotal voters' likelihood of voting to punish while holding the absolute number of group votes constant. To ensure participants understood the rules of the task, they answered an attention check and two comprehension checks—one about the rules of the Dictator Game, and one about which group outcome would occur for each possible voting distribution, given the participant's assigned voting threshold. All participants were required to complete this latter check correctly to proceed.

We measured participants' beliefs about the degree to which the target is deserving of punishment, and randomly assigned whether they reported these beliefs before or after learning about how their group members voted. Participants were asked, "To what extent do you believe Player 1 deserves to be punished (by subtracting 19 cents from his/her bonus)?" (1 = Player 1 does not deserve to be punished, 7 = Player 1 deserves to be punished). Participants either answered this question immediately after learning how the Dictator allocated the endowment and before learning how the rest of their group voted, or they answered it after both finding out how the Dictator's allocated the endowment and

about how their group voted. The answer choices, presentation order of group votes, and indication of what final group decisions would be reached if the participant chose a given answer were presented in counterbalanced order.

1.8.5 Study 4

This study used a 2 (voting decision outcome: punish vs. hung jury, within-subjects) x 2 (voting threshold: majority vs. unanimity, between-subjects) x 10 (Vignette: 10 unique scenarios, within-subjects) design to test whether participants judge pivotal voters as differently responsible for conclusive and inconclusive jury outcomes.

Participants and Procedure

We requested 570 participants on Amazon's Mechanical Turk, for a preregistered online study in return for $1.20. We sought a final sample of 400 participants and expected 30% of participants to fail our comprehension and attention checks. This process returned 576 participants, 14.1% of whom failed these checks, thus yielding a final sample of 505 participants ($M_{age} = 41.67$; $SD_{age} = 13.17$; 56.24% female).

Participants read 10 scenarios in which harm was done to someone and a jury with 12 members was deciding whether or not to punish a defendant involved in the event (these scenarios were drawn from Multistate Bar Exam ["the Bar"] preparation materials). For example, one scenario involves a jury deciding whether to send a doctor to jail for one year for negligence because, while dining at a restaurant, he did not help a person who was choking and subsequently suffered severe and preventable brain damage. Moreover, they learned that a minimum number of votes would be required for each jury to render a verdict (of whether or not to punish) and that failing to reach this minimum voting threshold would result in a hung jury—meaning that the verdict would be left to an entirely different jury.

We randomly assigned participants to one of two between-subjects voting threshold

conditions: Majority or Unanimity. In the Majority condition, all juries in a given participant's scenarios would need at least 7 out of 12 jurors to agree in order to render a verdict. In the Unanimity condition, all juries would instead need 12 out of 12 jurors to agree in order to render a verdict.

Participants read about a focal juror in each scenario who held a pivotal vote—meaning this juror held the deciding vote between either punishing the defendant or having the jury hang. The focal juror was always the last of the 12 jurors to vote on a jury that was one vote away from reaching a verdict. In the Majority condition, the current vote (prior to the pivotal juror casting any votes) was always 6 votes to punish versus 5 votes not to punish; in the Unanimity condition, the current vote was always 11 votes to punish versus 0 votes not to punish.

For each of the 10 scenarios, we randomized (within subjects) the voting decision outcome by having the pivotal juror either vote to punish—resulting in a punishment verdict—or vote not to punish—resulting in a hung jury. To confirm that participants understood the task and were paying attention, they completed an attention check and a comprehension check. We also randomized the order in which the scenarios were presented, and also randomized the names of the pivotal jurors from a pool of the 50 most common male and female names in the U.S. over the last 100 years, according to the Social Security Administration.

Our dependent variable of interest was participants' judgment of pivotal jurors' responsibility for the voting decision outcome. For each scenario, we collected three 'responsibility' measures (in randomized order), asking participants about the extent to which participants felt the pivotal juror was responsible for the group outcome (1 = not responsible at all, 7 = completely responsible), accountable for the group outcome (1 = not accountable at all, 7 = completely accountable), and causal of the group outcome (1 = did not cause it at all, 7 = completely caused it). Lastly, participants reported demographic information and whether or not they experienced technical issues.

Tables

Table 1.1

Regression models showing effects of pivotal voter status and conformity on the probability of voting for punishment and beliefs about deservingness of punishment

This table depicts three models. Model 1 is a logistic regression model with voting to punish the target as the dependent variable. Model 2 is an OLS model with ex-ante beliefs about targets' deservingness of punishment as the dependent variable and pivotal voter status as the independent variable. Model 3 is an OLS model with ex-post beliefs about targets' deservingness of punishment as the dependent variable and pivotal voter status and other group members' votes (to punish) as independent variables. Ex post and ex ante are defined relative to learning the social information (i.e., other peoples' votes). We find that being pivotal increases the likelihood of voting to punish. Pivotal voters also update their beliefs correspondingly—being pivotal increases the degree to which targets seem to deserve punishment. Statistical significance is denoted by *, **, and *** for $p < .05$, $p < .01$, and $p < .001$, respectively.

	Dependent variable:		
	Voted to punish	Ex ante belief	Ex post belief
	(1)	(2)	(3)
Pivotal Voter	0.465***	-0.065	0.487*
	(0.118)	(0.200)	(0.201)
Assigned Votes (for Punishment)	0.055	-0.002	-0.022
	(0.062)	(0.105)	(0.103)
Constant	-0.599	3.515***	3.894***
	(0.556)	(0.950)	(0.935)
Observations	1,633	815	818
R^2	–	-0.002	0.006
Note:			*p<0.05; **p<0.01; ***p<0.001

CHAPTER 2

PREDICTABLY BAD INVESTMENTS: EVIDENCE FROM VENTURE CAPITALISTS

2.1 Introduction

The most influential companies of today all began as startups in need of capital. Investors looking to capitalize on the next generation of superstar firms want to do two things: fund startups that will eventually be superstars and avoid lemons. In other words they face a prediction problem (Kleinberg et al. 2015). Previous work has explored what influences investor choices[1] as well as descriptive and causal evidence on the financial performance of investor portfolios[2]. However, little work has been able to tie the two together to identify systematic choices while quantifying their impact on performance. For example, in recent work, Zhang (2020) studies investor discrimination through field experiments, but it isn't clear whether this behavior is costly to the investors. This paper provides evidence that approximately half of VC investments are predictably bad relative to several outside options thus costing firms over $900 million.

Consistent with the predictive nature of investor decisions, I take a machine learning approach to evaluating the quality of their decisions. I first train an algortihm based on early information about each firm to predict the late-stage exit of the firm. The central eeconometric challenge with this exercise is the selective label problem (Kleinberg et al. 2018a)— we see the performance of firms that received investment, but we don't see the

1. Many related things are known in the venture capital space. Important work includes Gompers, Kaplan, and Mukharlyamov (2015) who surveyed 79 private equity investors with combined AUM of over $750B about their practices in firm valuation, capital structure, governance, and value creation. An incomplete list of other factors that have been explored includes product-market fit (Hellmann and Puri 2000), syndication strength (Lerner 1994), CEO personality characteristics ((Kaplan and Sorensen, 2017), Kaplan, Klebanov, and Sorensen 2012), pitch delivery (Hu and Ma 2021), and perceived passion (Gompers et al. 2020).

2. For example, the effect of VC monitoring on performance (Hellmann and Puri 2002; Lerner 1995; Kaplan and Stromberg 2001, Kaplan and Strömberg 2004).

performance that would have been realized by firms that didn't receive investment, if they had in fact received investments. I leverage the fact that the problem is one-sided and focus my evaluation on the firms that received investments. Were those investments good predictions or false positives? Importantly, I take the ex ante view in answering these questions.

Using data from Pitchbook on the universe of startups accepted into the top accelerator programs, I first document new descriptive evidence on the venture capital (VC) industry, documenting major accelerators, investors, and markets that startups participate in. Next, I find that a standard off-the-shelf algorithm can strongly discern between startups that are more or less likely to have exits. That an algorithm can detect a statistically reliable pattern opens the scope for investors to have returns not by luck, but by selecting on quality. Indeed, I observe a nominal return of 79% among VCs, which outperforms the S&P500 by 15 percentage points over the same period. To evaluate room for further outperformance, I use the algorithm to build counterfactual portfolios forgone by investors by pruning predictably bad investments and opting for a more standard outside option (a public equity or bond instrument). I find that despite the fact VCs outperform the market on average, the returns are driven by the top half of the predicted quality distribution. By dropping the bottom half of investments and instead investing in the market, returns would have increased by 7 to 41 percentage points. This qualitative finding is robust to a set of outside options (stock market or bond market). Together the results suggest that there is significant room to improve how venture capitalists select into investments.

Finally, I investigate why investors consistently invest in companies that are predictably doomed to fail. I provide evidence for one key bias: overweighting of human capital information. I find that investors overweight the importance of founder characteristics, especially for lower-quality firms. I explore several mechanisms underlying these poor investments. I assume throughout the main analyses that investors and their principals attempt to

maximize returns, making the deviations I document persistent mistakes. I discuss relaxations of this assumption in Section 2.6 and explore alternative interpretations.

Notably, my main empirical analyses are closely related to independent but contemporaneous work from Lyonnet and Stern (2022) (LS) as both closely follow Mullainathan and Obermeyer (2022). While my qualitative conclusions are similar to those of LS, I note several key differences. First, I emphasize that investors as well as the researchers studying them have focused too much on the jockey and not enough on the horse; instead of documenting specific biases about how founder information is prioritized, I find that founder information as a category is of second order importance, relative to information on the business itself (i.e., the horse). Second, I observe the actual investments made by investors, which allows me to calculate actual returns and estimate the economic magnitude of returns left on the table. Third, I restrict attention to the consideration set of venture capitalists by focusing on startups that participate in top accelerators, whereas LS considers a highly inclusive definition of startups, studying (a representative sample of) *all* new companies in France. Finally, I study startups competing for funding in the US VC landscape which is significantly more developed and competitive than that of France. Showing that the results hold in the US underscores the persistence and importance of my documented biases in highly competitive markets.

This work contributes to several literatures. First, a growing literature on **behavioral industrial organization** has assessed whether biases can persist in equilibrium in markets with large stakes, sophistication, and expertise. (List 2003, List 2006, Benz and Meier 2008, Strulov-Shlain 2021). In particular, I emphasize deviations that behavioral firms are making from a benchmark of profit maximization as in DellaVigna and Gentzkow (2019), Bloom and Van Reenen (2007), Di Maggio, Ratnadiwakara, and Carmichael (2022) and Hortaçsu and Puller (2008). I show that highly costly behavior can persist even among firms with strong financial incentives. Startup investors appear to be susceptible to the same attentional and

information processing biases observed in consumers by both psychologists and economists. (Gabaix 2019, Howell 2020, Kahneman 2011, Enke 2020)

Methodologically this paper further demonstrates the value of **machine learning to answer economic questions**. My results suggest that venture capital investing is yet another setting that can be improved by the human+machine paradigm. (e.g., Dawes, Faust, and Meehl 1989, Kleinberg et al. 2018a) In addition to being practically useful to investors, this approach has generated useful research insights across several fields in economics, such as health, law and economics, labor, and development, but research on venture investing has yet to fully leverage these tools.[3] (Mullainathan and Obermeyer 2019, Kleinberg et al. 2018b, Jean et al. 2016, Bansak et al. 2018). In addition, since the most influential data to the model is text, my approach adds to a newer, methodological literature on "text as data". (Gentzkow, Kelly, and Taddy 2019, Gentzkow and Shapiro 2010, Erel et al. 2021, Ke, Kelly, and Xiu 2019, Kogan et al. 2009).

Finally, the source, persistence, and opportunities for future returns has long been a fundamental question in **finance**. (Fama 1965) Of particular interest is whether returns can persist, which can be framed as distinguishing between investors getting the prediction problem right or just getting lucky. (e.g., Kaplan and Schoar 2005, Nanda, Samila, and Sorenson 2020) My results suggest that, at least in private markets, there is significant scope for persistent performance, which would be realized by investors if they relied less on the founders, informing the "horse versus jockey" debate. (Kaplan, Sensoy, and Strömberg 2009, Zingales 2000, Gompers et al. 2020, Rajan 2012)

The rest of this paper proceeds as follows. Section 2.2 provides a framework for assessing the startup landscape. Section 2.3 describes the data and how I use it in an ML framework. Section 2.4 documents new empirical facts about the startup lifecycle. Section 2.5 evaluates

3. Previous work has used machine learning to predict a variety of outcomes (Ghassemi, Song, and Alhanai 2020; Ying, Wua, and Lia 2021, Bento 2018, Żbikowski and Antosiuk 2021, Krishna, Agrawal, and Choudhary 2016, Hu and Ma 2021) but none of these have taken into consideration what information was plausibly available to investors and then used that to evaluate investor actual choices.

realized and forgone returns. Section 2.6 explores the underlying causes and how these mistakes persist. Section 2.7 concludes.

2.2 Framework

2.2.1 Startup Lifecycle

To guide the empirical analysis, I summarize the startup lifecycle with a simple two-period game (depicted in figure 2.1):

1. **Prediction policy problem**: Early-stage investor evaluates the startup and decides whether to invest

2. **Outcome realization**: Startup realizes its fate. A set of late stage investor decides to invest giving the early stage investor an opportunity to liquidate the investment and realize gains, if any.

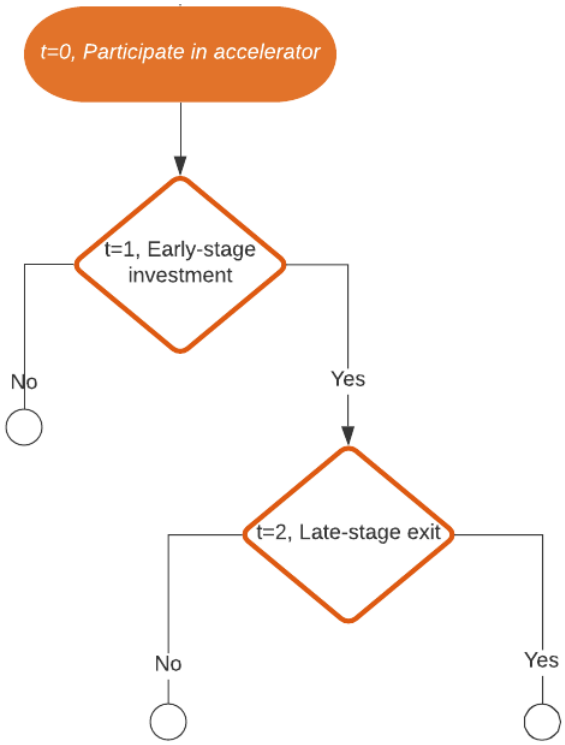

Figure 2.1. Timeline of events. t=0 defines the sample; All companies that participated in any of the top 100 US incubators or accelerators before 2019 are followed through the end of 2020. Two key outcomes are measured: early-stage investment (t=1) and late-stage exit (t=2). Only information available as of the incubator stage (t=0) is used to predict the two outcomes.

Formally, startups, indexed by i, are evaluated by investors, indexed by j, in period 1 and have a late stage exit in period 2. Further, let there be characteristics of the startup X (which are observed by both the investor and the researcher) and Z (which are observed only by the investor); both sets of characteristics may change over time. The key choice for the investor is to choose a level of funding $f \geq 0$. Then, in period 2, the startup realizes an exit with probability p, which obeys

$$p_i = \theta_{i,t=2} + T_{i,t=2} \tag{2.1}$$

where $\theta_{it} = \theta(X_{it}, Z_{it}) \geq 0$ is a measure of the intrinsic quality of the firm and $T_{it} = T(X_{it}, Z_{it}|f_{t=1}) \geq 0$ captures the treatment effect of the funding f the startup received in

the first period.[4]

As is standard, the investor is unable to directly observe θ and T. However, in this setting the difficulty for the investor is magnified because θ is a function of $(X, Z)_{i,t=2}$ which is potentially different than the information available to the investor at the time of investment $(X, Z)_{i,t=1}$.[5]

2.2.2 Investor Choice and Payoff

For a given level of funding f, the firm's expected return is

$$r(\theta, T) = pM$$

where M is the multiple on invested capital the investor would receive in the event of a successful exit. Return-maximizing investors then face a straightforward investment decision: invest in firm i if $r(\theta, T) \geq \rho$ where ρ is the expected return from some outside option. Investors then need to form some prediction of r,

$$\hat{r}^h(X, Z, b)$$

which is a function of the information (X,Z) available to the investor as well as a vector of potential behavioral distortions b, including non standard preferences, belief distortions, or attentional constraints.

2.2.3 Evaluating Investor Choice

An alternative to forming human predictions of r is to build an algorithm that can generate predictions $\hat{r}^a(X)$. This approach has several benefits to the researcher. First, it employs

4. I assume T is nonnegative, but make no assumption about its shape.

5. The calendar time that typically separates period 1 and 2 is five years

an analysis that the firm could have and may have done in order to improve investments. Second, such an algorithm has the potential to proxy complex prediction rules made by humans. Third, the algorithm is not subject to b. That is, there is no opportunity for human distortions to influence predictions.

However, there are also two key drawbacks to this approach as discussed in (Kleinberg et al. 2015):

1. Selective labels: don't know counterfactuals since neither T nor θ are known
2. Selection on unobservables: Z is unobserved and potentially influencing exits and investments.

To address both problems I form predictions on those firms that received investment and assess which of those should have been avoided because $\hat{r}^a(X) < \rho$. In other words, my approach allows me to identify false positives; no further statements can be made about 'missed opportunities' without further assumptions, which I leave to future research.

To see how this approach addresses the central econometric challenge, note the relationship between p from equation (1) and the data we observe. Regardless of whether the startup received investment, we can define a binary realization Y of exiting. Then for any cell X, $\bar{Y}$ is an unbiased estimator of p iff all startups in the set received funding. For any cell X, $\bar{Y}$ is an unbiased estimator of θ iff no startups in the set received funding. In other words, without directly estimating the treatment effect, by construction we know that treated firms have realized their treatment effect T, whereas unfunded firms are relying strictly on their inherent type θ. This is an instantiation of the selective label problem with a mixed prediction problem (Kleinberg et al. (2015)). As a result, the algorithm generating $\hat{r}^a$ can only credibly be estimated and evaluated on firms that did in fact receive funding.

With this framework in mind I will take systematic decisions to invest when $\hat{r}^a < \rho$ as evidence of a predictably bad investment.

2.3 Data and Methods

The sole data source for this paper is Pitchbook Data. Pitchbook is a subscription data provider widely used by investors for information on deals, companies, and other investors in private capital markets. This data is generally useful to academics because it provides a thorough view of the private market, but it is especially useful for this paper because I will rely on data readily available to the investors I study, emphasizing the extent to which investors misuse available (albeit costly) information. There are three categories of information that I synthesize from the Pitchbook data: *finances (e.g., Revenue, EBITDA, total capital raised), founder information (e.g., educational background and previous experience)*, and *company/product description*. I describe the construction of the sample and relevant variables below.

2.3.1 Defining the sample

I begin by identifying every startup that participated in any of the top 100 accelerator and incubator programs between 2009 and 2016.[6] For each of those 16,054 firms, I then construct a dataset of all equity deals known to Pitchbook within the first five years of completing their accelerator program. One key motivation for pre-defining a set of firms of interest and then following their valuations over time is to avoid any survivorship bias which would severely limit the interpretation of any results.

Because my primary empirical strategy relies on prediction techniques, it is important to distinguish between the subsample used for training the algorithm and the subsample used to evaluate it. All models are trained on companies that completed accelerators before 2014. To avoid overfitting concerns, all results and figures are based exclusively on the hold out

6. Using the Pitchbook data, investors who are tagged to the Accelerator/Incubator investor type produces about 7,000 investors which I sort by the number of Accelerator/Incubator DEALS they have participated on, then pulled the top 100. Many of the companies will be familiar to the reader and include Airbnb, Doordash, Stripe, Dropbox, Coinbase, Instacart, and Uber.

data that covers 2014-2016. There are several advantages to taking a time-based sample splitting approach. First, it allows me to mechanically avoid any hindsight bias. Second, the empirical exercise simulates what forward-looking investors could have done in real time. And third, the analysis implicitly tests for the stability of predictiveness over time.

Notably there are two reasons why predictive quality may suffer when moving from the training set to the test set. The first is the traditional overfitting concern. The second is changes in the covariance structure between X and Y. These two concerns make the current exercise a conservative test of the predictability of exits and returns. But ultimately, the goal is not to reach a certain threshold of predictive accuracy per se. Rather, my central goal is to assess and improve investor decision quality and by building an algorithm based on information that investors had access to, I can simulate different human+algorithm policies that were implementable.

2.3.2 Raw predictors

For each funding deal, Pitchbook provides numerical data on the financials of the firm and qualitative information about the CEO, *at the time of the deal*. This time-specific nature of the data is central because my primary empirical strategy will make predictions of late-stage outcomes exclusively using information available to post-incubator investors. Accordingly, nearly all predictors will come from the financials and deal information as of participation in the incubator. This includes numerical information such as Revenue and EBITDA, but also includes the terms of participation in the incubator such as the level of investment and the equity stake that the incubator will take in the startup. Collectively I refer to these variables as *financial information*. The data also contain text fields that include biographical and educational information about the CEO, which I will refer to collectively as founder information.[7]

7. To make use of the rich information embedded in the text fields, I employ a step of borrowing techniques from natural language processing in the machine learning literature. Specifically I vectorize the text fields

The one data point that is not specific to the incubator deal is the company description. Pitchbook does not store or provide a time-varying description of the company. Instead, for each startup I have a current description of the firm, its product, and activities, independent of its health or status. I use this information in my analyses as a way to extract information about the company's product and thus refer to this field as product information. A key assumption in this approach is that the descriptions found in Pitchbook do not evolve as a function of early-stage funding or late-stage success which would mechanically make the descriptions predictive. To the best of my knowledge the company descriptions available to me today are not systematically different from the descriptions available to early-stage investors when they would have been engaged in due diligence. Further, all results are robust to the exclusion of the product information.

Tables 2.1 and 2.2 provide summaries of the sample and the key predictors.

2.3.3 Outcomes

There are two categories of outcomes that will be central to the main results: post-incubator early-stage investment and late-stage exit.

I define early-stage investment as any equity deal in the Pitchbook data within two years of incubator completion that is categorized as "Series A", "Series B", "Seed Round", or "Angel (Individual)" in the Deal Type. For much of the analysis I will collectively refer to investors at this stage as early-stage investors though some analyses will separately identify investors in Series A or B rounds as venture capitalists and other investors as Angel/Seed. 34% of the startups receive early-stage investments.

I then define late-stage exit by assessing whether a startup has any of three eligible late-stage deals/transactions: initial public offering (IPO), merger or acquisition (MA), or any

using 1 and 2 grams and reweight using a TF-IDF algorithm, all within the training data. The resulting transformations are then applied to the test set. See Gentzkow, Kelly, and Taddy 2019 for a recent review on empirical research relying on transformations of text data.

funding round that is categorized by Pitchbook as Series C or later (C+) within five years of accelerator completion. This is a fairly liberal definition of exit because it neither requires a liquidation of a stake nor a positive return. Justifying an IPO as an exit is straightforward as the startup is publicly sold and the early investors have a clear opportunity to liquidate and profit. M&A transactions are murkier. While the investor may have the opportunity to sell their shares, it is not necessarily expected that there will be a positive return from the transaction. With an eye towards this I will incorporate in my main analyses measures of financial return, usually multiple on invested capital (MOIC), along with binary indicators of success in order to avoid misinterpreting situations where the early-stage investor successfully sells a stake in a MA transaction albeit at a steep economic loss. C+ transactions present the opposite problem. I limit the eligible C+ transactions to those late-stage rounds that are classified by Pitchbook as "Up" rounds to avoid including deals where the the startup has raised more money, though at a lower valuation, but I cannot observe whether the investor has the opportunity to liquidate their ownership and therefore realize any gains from the transaction. Nonetheless, recent evidence suggests that for the Series C round and later, there is ample opportunity to sell shares on a secondary market. Using this three-pronged definition, 3% of the overall sample has a successful exit and 9% of the startups that receive early-stage investment have a successful exit.

In the sample there are 5,440 early-stage investments totalling over $9.3B dollars and 497 late stage transactions totalling over $105B in company value.

2.3.4 Missingness

Many of the predictors are missing not at random. If my objective was to provide estimates for the influence of any particular variable on an outcome, this missingness would be a major vulnerability for interpreting the results. But since my only goal is to find a predictive signal, the absence of information is potentially a valuable signal in itself. Accordingly, any

observation with a missing value for any variable is coded with a filler value so it can be flexibly handled in the algorithm learning stage.

There is also missingness in some of the outcomes which needs to be handled differently. If a startup does not have a Pitchbook-documented late-stage exit by the end of the five-year horizon, then I assume the value of the firm is zero to reflect the realized valuation by the investors and their principals. If a startup has a late-stage exit but the valuation is not known, I also assume it is equal to zero. If the initial stake of the early-stage investor is unknown then I assume they purchased 30% of the company, which is standard for Series A investors.

2.3.5 Constructing an algorithm

A traditional econometric approach might consist of estimating an OLS model and then using that model to form $\hat{y}$ predictions. This traditional approach will not suffice in this setting for two reasons. First, most of the data is text and OLS would require taking a strong stance on how to code the data. Second, and perhaps more importantly, instead of imposing linearity in the relationships, more flexible algorithms allow for arbitrary relationships and interactions between all the variables and take a data-driven approach to assessing which interactions and relationships are reliable. To that end, I train a model using XGBoost on numerical transformations of the text data combined with the native numerical data.

2.3.6 Calculating performance and returns

The primary metric that I use to evaluate returns is "multiple on invested capital" (MOIC), which is common in both academic research and industry. It's calculation is simple:

$$M = \frac{\text{Value of investment}}{\text{Initial investment level}}$$

I describe the specific calculations and considerations for each term in turn.

Initial investment level

For analyses that include MOIC, I restrict attention to those companies where Pitchbook provides data on the total invested equity in the early funding round following accelerator participation. No adjustments or imputations are imposed on these initial investments.

Value of investment

The basic equation for calculating the value of the investment is

$$\text{Value of investment} = \text{Value of firm}_{t=2} * \text{Initial ownership } \% * \text{Dilution factor}$$

Several of these quantities are not provided in the data and therefore need to be imputed or assumed. First, the valuation of the firm at t=2 in model time is defined as the latest valuation associated with a "late-stage" round by the end of 2020. If no such rounds exist then the company valuation is set to zero. If such a round exists but the valuation is not provided by Pitchbook, I set it to zero. Next, Pitchbook typically provides the initial ownership stake of the early stage investors. If it does not then I assume the early stake is 30%. Finally, I assume a constant dilution factor of .75.

2.4 Descriptive Statistics on Startups, Investors, and Performance

2.4.1 Accelerators by the numbers

Accelerators are seen as a launch pad for many startups. This section documents new data on how accelerator cohorts have performed from 2009-2016. Table 2.5 provides several basic statistics for the top 5 accelerators in the Pitchbook data. The key statistic is the sum of the post-money valuations for firms launched from a given accelerator. The top two (Rocketspace and Y Combinator) support several common, though conflicting intuitions about startup investing. First, Rocketspace, while only investing in 3 companies in my data, is ranked at the top. It's sole non-zero investment is Uber. This supports the notion that one right-tail outlier can define the performance of the entire portfolio. The second investor in the table is Y Combinator which has a much broader scope, investing in over 400 companies over the course of my data. These two investors together account for over 70% of the market value of companies in my data, suggesting either a strong ability for these accelerators to discern on type θ or a strong treatment effect T of the community, mentoring and advising that they aim to provide.

2.4.2 Investors and performance

Table 2.4 presents data similar to that in Table 2.5, instead focusing on the investors who seek out alum from the accelerators. The names of the investors in the top firms will be familiar to practitioners: Benchmark, Seqouia, and Accel. These firms again appear to have some ability to discern since they make up relatively small shares of total investment, but comprise outsized shares of market value of portfolio companies. For example, Sequoia Capital only makes up 1% of invested early-stage equity in my data, but its portfolio companies make up 6% of market value in my data.

2.4.3 Markets

Finally, Table 2.6 presents the markets that are most represented among accelerator participants. Column "% of All", for example, shows the distribution of markets among all accelerator participants. I interpret differences in the columns as measures of how much certain markets are preferred at different stages of the startup lifecycle. For example, 8.5% of the full sample operates in the "Application Software" market. Whereas only 5.9% of the firms receiving early stage investments are in the Application Software market. This crudely suggests that Application Software is underrepresented in early stage investments and that investors appear to slightly prefer the opportunities in other markets to those in Application Software.

Haven given a broad, descriptive overview of an important collection of startups and the investors who fund them, I now turn to evaluating the quality of their decisions.

2.5 Evaluating Exits and Returns

2.5.1 Startup success is predictable

Algorithm can discern success rates

I begin by showing a monotonic relationship between the predicted success of startups in the hold-out set and their observed success, thus demonstrating that success is predictable early in the life cycle of a company.[8] The algorithm has found signal. That startup success is predictable at all creates scope for savvy investors to have persistent returns which remains an outstanding question in the literature.

Figure 2.2 compares the predicted success measure to the outcome for which it is trained: a binary indicator for IPO, MA, or C+ funding round. The x-axis rank orders startups by

8. Given the selective label problem I restrict attention to the 5,440 startups that received investments.

their predicted success and the y-axis shows the fraction of firms that achieve an exit. Those firms deemed poor investments by the algorithm successfully attain exits at a rate of 5%, whereas the top predicted investments exit at a rate X times higher, despite the fact that I test the model on companies it has never seen, thus avoiding overfitting concerns. What's especially striking is the implication that what was predictive from 2009-2014 (what the model was trained on) remains predictive on later data (what the model is tested on), allaying concerns about distribution shifts over time.

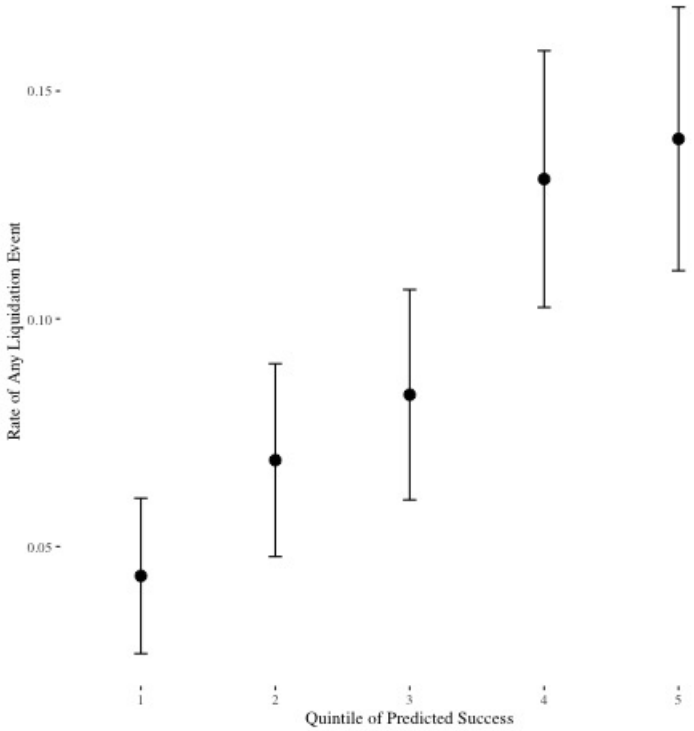

Figure 2.2. Late-stage outcomes. Relative frequency of exit shown by quintile of predicted success produced by the ML predictions. Exit is defined as an IPO, acquisition, or Series D or later transaction. Companies with no late-stage transactions are valued at $0. Sample is subsetted to those companies that received early-stage VC investments. Predicted success deciles are defined within the subsample.

That the model can predict the outcome it was trained to predict is informative for evaluating the model itself, but does not speak directly to the investors' decision problem. Investors want to maximize the dollar value of the return, not simply the probability of an

exit. I improve the analysis by showing the distribution of valuations in each bin of predicted success. Figure 2.3 shows two striking empirical facts. First, companies with higher predicted success are more likely to have positive valuations. Second, the split between positive, low-value and positive high-valuations shifts dramatically to the right as one gets to higher levels of predicted success.[9]

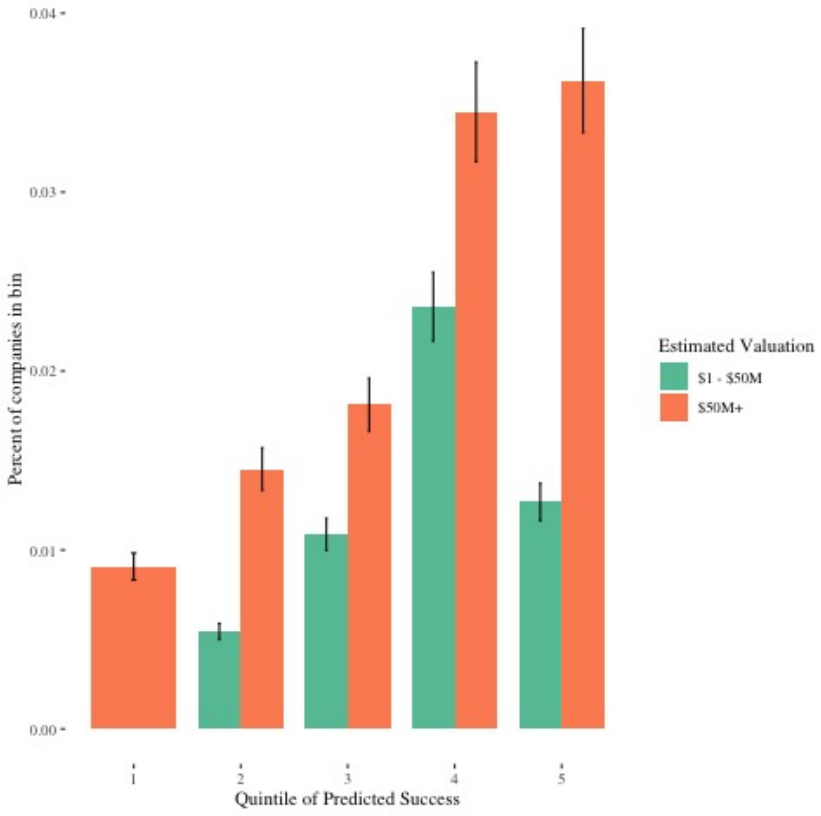

Figure 2.3. Distribution of valuation outcomes. Relative frequency of each valuation bin shown by decile of predicted success produced by the ML predictions. Companies with no late-stage transactions are valued at $0 and are thus not shown. Sample is subsetted to those companies that received early-stage VC investments. Predicted success deciles are defined within the subsample.

9. For the reader interested in more traditional measures of predictive quality from the machine learning literature, the model's ROC curve is shown in Figure 2.4.

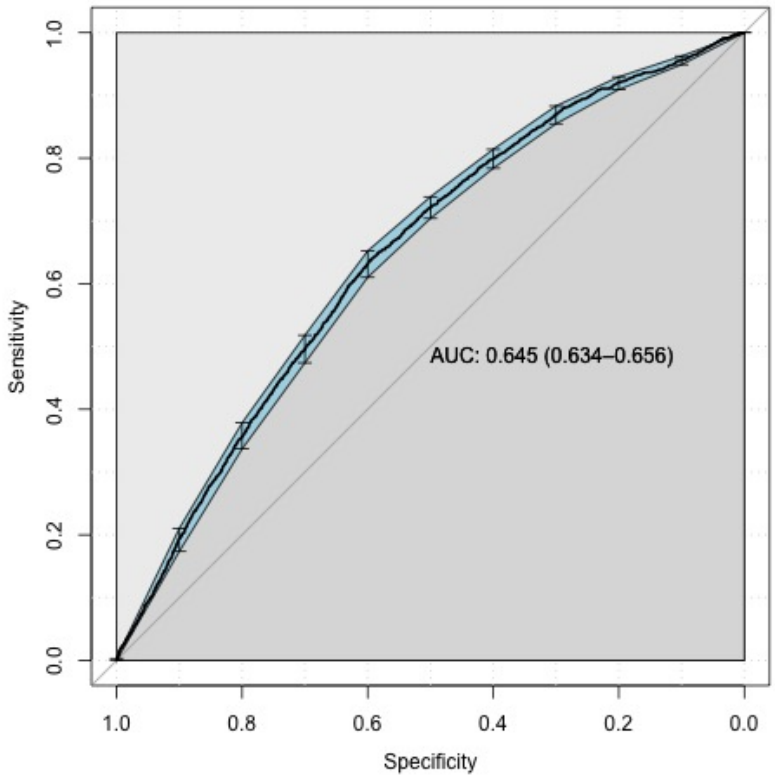

Figure 2.4. ROC-Curve. This graph shows the Receiver Operating Characteristic Curve for the algorithm trained to predict late-stage exit. The area under the curve (AUC) is a common measure of the predictive quality of the algorithm. An AUC of 0.5 is equivalent to random guessing and improvement over 0.5 indicates an ability to predict with some level of accuracy.

What predicts success?

What is the algorithm using to predict exit so successfully? Discerning how the inputs to an algorithm relate to its predictions is notoriously difficult. For ease of interpretation, I project the algorithm's predictions onto common variables discussed in the literature, including level of funding, proxies for founder education quality, and gender. Table 2.7 reports the projections onto the model's raw predictions for all startups; Table 2.8 reports the projections onto the model prediction percentiles among the firms that did receive early investments.

The projections are consistent with the existing work that suggests founder education

and prior investments increase the probability of successfully exiting later in the startup lifecycle. However, I emphasize that nothing in these projections can or attempts to speak to a causal interpretation of X on Y. Instead they give some guidance on what is or could be used to form $\hat{Y}$.

On failure rates

Given how well the predicted success positively correlates with various well-observed measures of actual success, one might expect predicted success to negatively correlate with a notion of failure. In addition to the funding rounds data described in the Data section, Pitchbook also provides data on two events that indicate business failure: bankruptcy declaration and documented shut down. I show the relationship between failure and predicted success in Figure 2.5. Counterintuitively, the figure seems to show no clear relationship between the firm's predicted success and its probability of failing. If true, such a finding might be fatal to the validity of the model. However, given the high observability of the success outcomes and the high potential for mismeasurement and selection issues in the failure data, I interpret this contradiction as evidence that there are severe reporting biases in failure. It appears that those firms that had good chances of success are more likely to formally shutdown their business *in an observable way*. I therefore caution future research against interpreting the failure data as a source of truth without accounting for endogenous reporting.

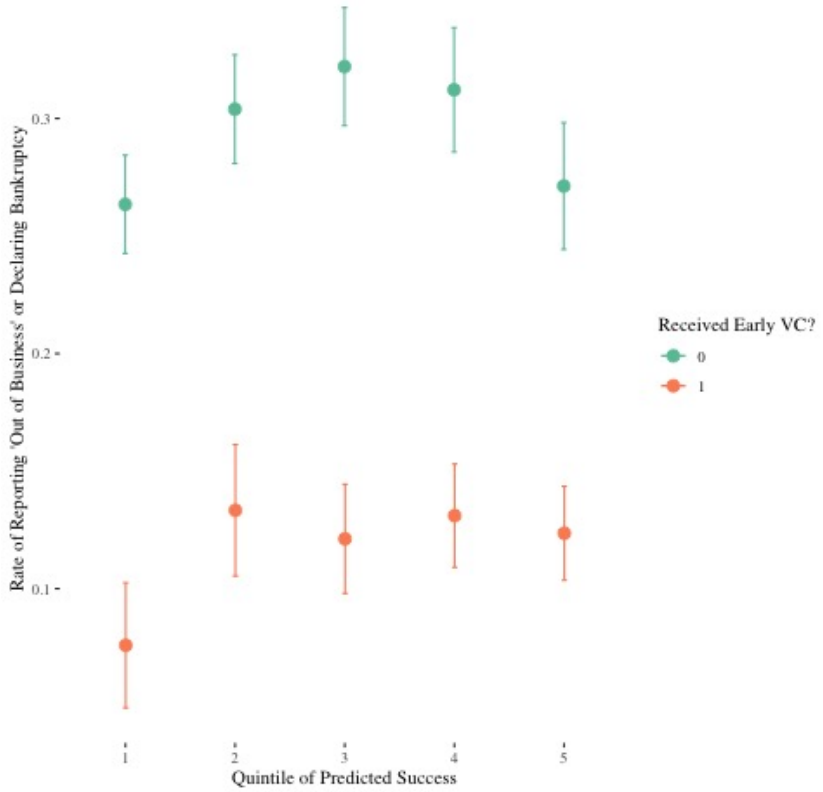

Figure 2.5. Documented failures. Proportion of firms that have a documented shutdown or bankruptcy filing shown by quintile of predicted success.

2.5.2 Assessing returns

In the following subsections I will evaluate returns through several strategies

Observed returns

The exit rate is higher in the firms predicted to do well by the algorithm, but the returns aren't as straightforward. Returns do not vary consistently with predictions as seen in Figure 2.6, which implies contracts and investment terms matter and possibly contribute to returns.

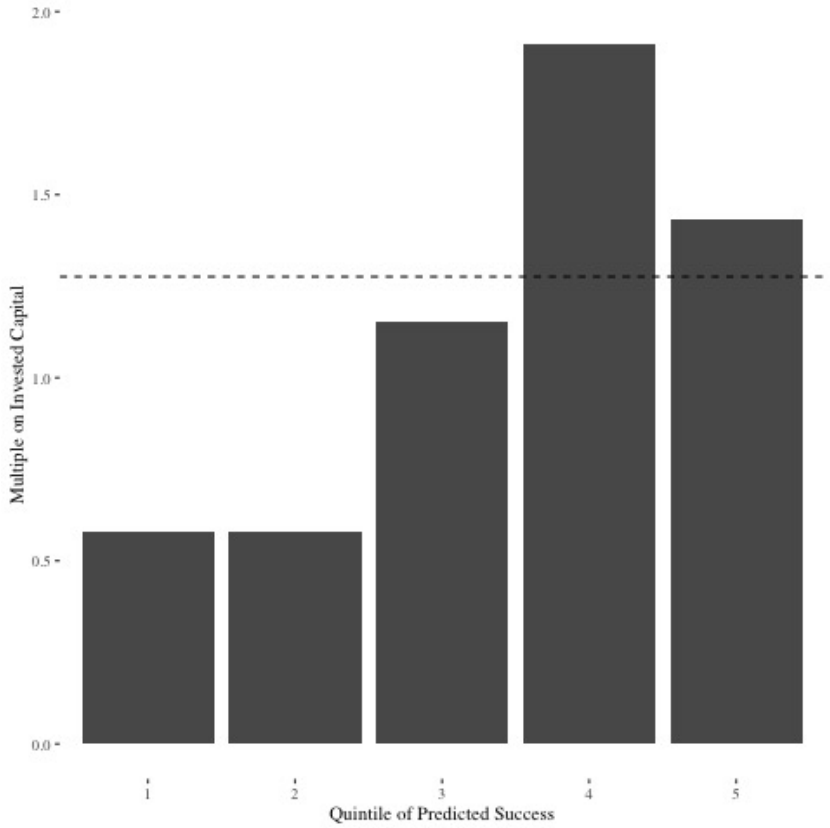

Figure 2.6. Returns by predicted quality. This graph shows the multiple on invested capital (MOIC) for firms in each quintile of predicted success. MOIC is defined by the ratio between the five-year valuation of the stake in the company and the level of the initial investment for that stake.

One way to explore the quality of the investments is to see if they differ by investor type. To do this I decompose figure 2.6 to compare individual investors (Angels) and institutional investors (venture capitalists). These two groups differ in several important ways. One prominent difference is that Angels invest their own money whereas VCs primarily invest other people's money and thus have broader scope for agency problems.

Figure 2.7 shows that the individual investors overall have lower returns than institutional investors suggesting that the sophistication, access to investments, and incentive scheme of the VC is enough to overcome the potential agency problems.

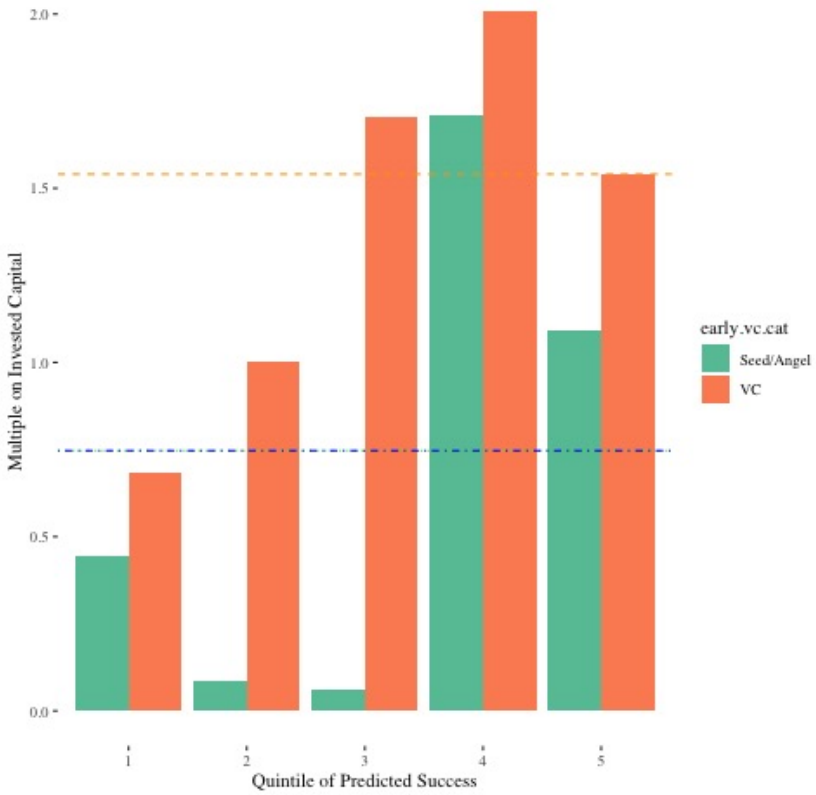

Figure 2.7. Returns by predicted quality and investor type. This graph shows the multiple on invested capital (MOIC) for firms in each quintile of predicted success separately for individual investors ("Seed/Angels") and institutional investors ("VC"). MOIC is defined by the ratio between the five-year valuation of the stake in the company and the level of the initial investment for that stake. The orange, dashed line is the overall MOIC for instituional investors and the blue dot-dash line is the overall MOIC for individual investors.

These results show that VCs outperform two common and intuitive benchmarks: the return of other investors and the average return of the outside option. However, as the framework in section 2 suggests, neither of these benchmarks is grounded in economic theory and therefore do not speak to the true optimal benchmark.

Knowing the true optimal benchmark, that is asserting the exact valuation and investment each investor should have for a given startup is not feasible in my data. The core challenge to such an exercise would be assessing counterfactual contract specifications, which requires knowing the function relating success to funding ($\frac{dT}{df}$). Without estimating inherent types and treatment effect functions, evaluating counterfactual investment contracts on the intensive

margin is not possible. However, evaluating counterfactuals on the extensive margin is feasible—what would've happened to f (the observed investment) if it wasn't invested in a given startup and instead was placed in an outside option.

The following subsection explores how well firms are performing relative to something closer to a second-best benchmark by comparing *each investment* against an outside option to construct a return-maximizing portfolio.

Forgone returns: Counterfactual bond portfolio

In the first set of counterfactuals I consider, the VC chooses between an observed investment contract or a hypothetical outside option. In particular, I consider a hypothetical bond that pays 8% simple interest.

One thought experiment to demonstrate how investors might improve their performance with an algorithm is to have the algorithm filter out the bottom k% of human startup selections and replace them with the outside option. Figure 2.8 graphs the results of such an analysis. Each point on the x-axis indicates a different scenario in which the aggregate investor invests only in the top 1-k% of companies according to predicted success. For example, the left most point is the return realized if the investor invests in the top 100 percent (i.e., all investments), thus yielding the exit rate that was actually observed. Then moving to the right, we observe how the VC portfolio multiple would change as a result of having an increasingly selective k, all the way up to only investing in the top 1 percent of startups.

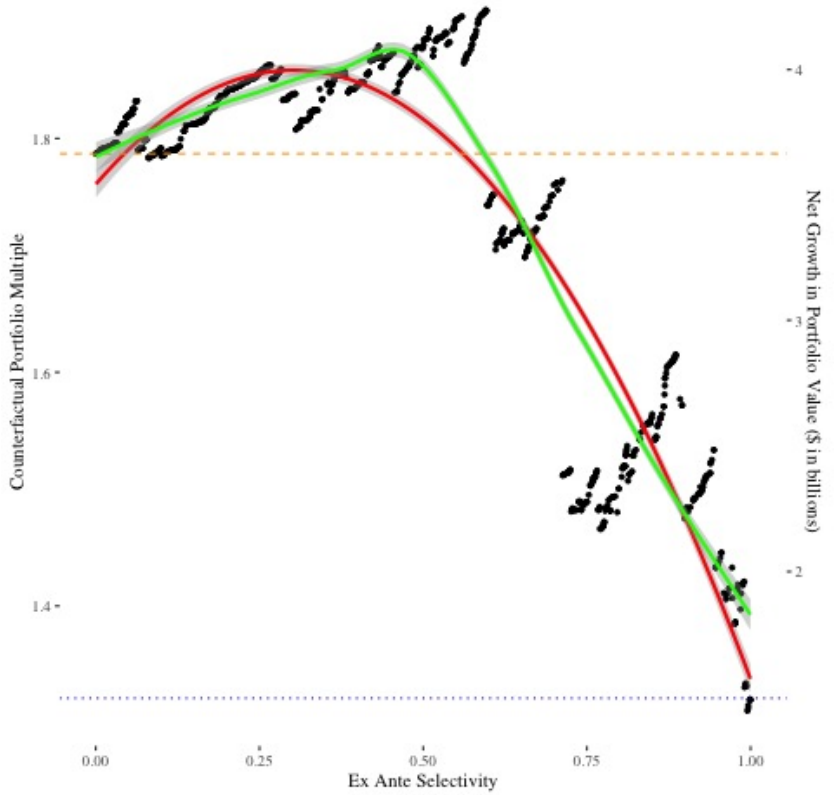

Figure 2.8. Counterfactual Payoff with Fewer Mistakes. Nominal return on investment that would have been realized by VCs had they had selected the outside option instead of investing in the bottom k percent of startups. Values above $1B are replaced with a $1B valuation. The outside option is assumed to be invested in a bond that pays 8% interest per year. The orange dashed line is the actual multiple on invested capital (MOIC). The blue dotted line is the MOIC the VC's would have realized if they instead invested all money in a 7% bond. Improvements above the orange dashed line indicate forgone returns that were predictable. A total of $6.25 billion in initial investments is captured in the graph, which provides the mapping between the two y axes.

Forgone returns: Counterfactual SP portolfio

Another approach is to consider a real outside option: the stock market. In Figure 2.9 I graph the resulting analysis. Instead of witholding investment as above, the investor invests the money in the S&P500 index. This exercise is very similar to the Public Market Equivalent (Kaplan and Schoar 2005), but differs in that I do not directly observe cash flows.

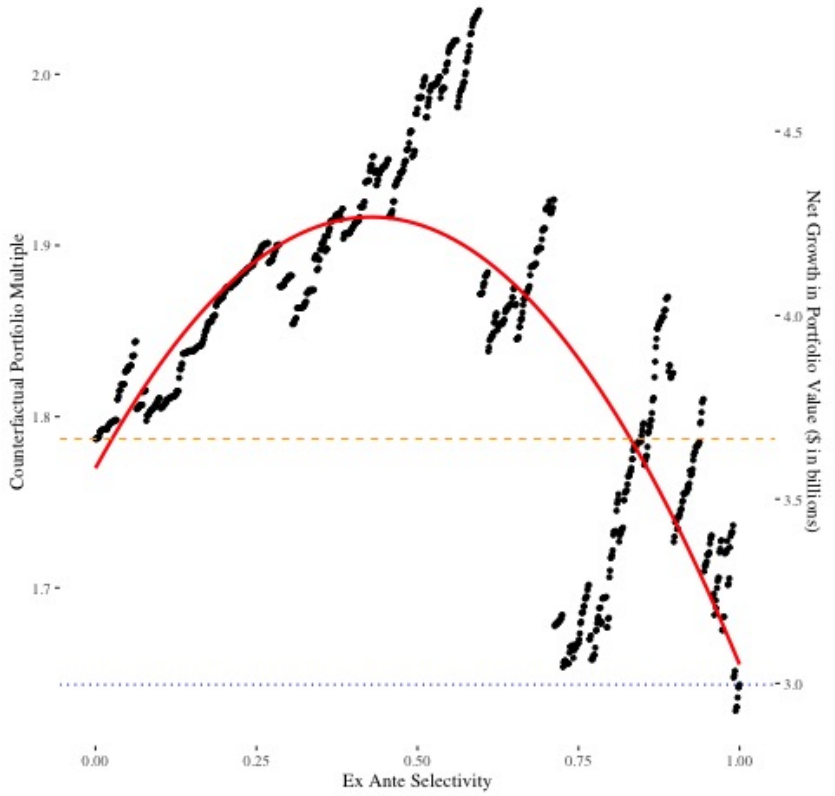

Figure 2.9. Counterfactual Payoff with S&P Outside Option. Nominal return on investment that would have been realized by VCs had they only invested in the top k percent of startups. Any money that is not invested in a startup is assumed to be invested in the S&P 500 index. The red dashed line is the actual multiple on invested capital (MOIC). The blue dotted line is the MOIC the VC's would have realized if they instead invested all money in the market. Improvements above the red dashed line indicate loss of alpha due to mistakes. A total of $6.25 billion in initial investments is captured in the graph, which provides the mapping between the two y axes.

In this analysis I find qualitatively similar results, though the returns to selectivity are higher since the return to the stock market option is higher than that of the hypothetical bond.

The two counterfactual portfolio analyses yield three takeaways. First, current VC returns are significantly higher than simple returns from less risky investments. Second, VCs could reliably drop up to 50% of their investments and improve returns by avoiding low-return startups. Third, VCs cannot trim investments all the way up to the top 1% because eventually they start swapping high quality investments for lower returns. There are at least

two interpretations of these empirical facts. First, is a literal and extreme interpretation: investors might do well to adopt a selection model close to that of CalTech, selecting only a few, highly promising startups. An alternative interpretation is that these analyses provide suggestive evidence that investors should be seeking out better deals terms. Under either interpretation I conclude that VC returns could be up to 1,000 basis points higher if investors avoided more predictably bad investments, on either the intensive or extensive margin.

Robustness

One prominent limitation in the approach above is the dependence of the analysis on a single ex post realization of the world. To address the concern that my results may not be robust to another draw from the superpopulation of portfolio outcomes under a given investment strategy, I conduct a bootstrapping procedure where I draw a sample of candidate startups (with replacement), implement a constant selectivity decision rule, repeat 2,000 times and report a distribution of statistics stemming from the procedure. Table 2.9 provides suggestive evidence that the hill-shaped relationship in Figure 2.9 is not due to chance. Though the estimates are noisy, the vast majority of the range, under any calculation, has the correct sign. Figures 2.10 and 2.11 provides further support by showing the distribution of MOIC that would have been realized under the current investment strategy versus a more (ex ante) selective one. This figure shows that the returns to switching to a more selective strategy were positive in over 95% of the bootstrapped samples.

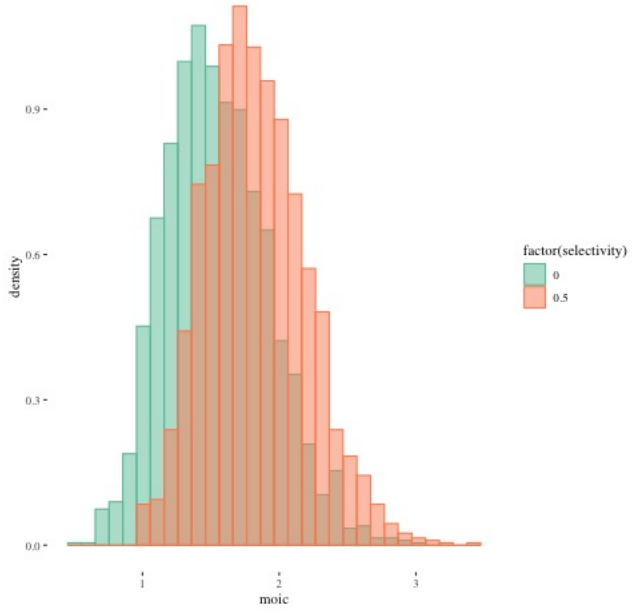

Figure 2.10. Bootstrapped estimates of MOIC. This graph quantifies the distribution of performance gains derived from a bootstrapping procedure. The green bars indicate the distribution of MOIC that is realized under the current investment strategy, while the orange bars indicate the distribution of returns under a strategy of dropping the lower half of startups (according to the ex ante quality measure) in favor of the S&P 500.

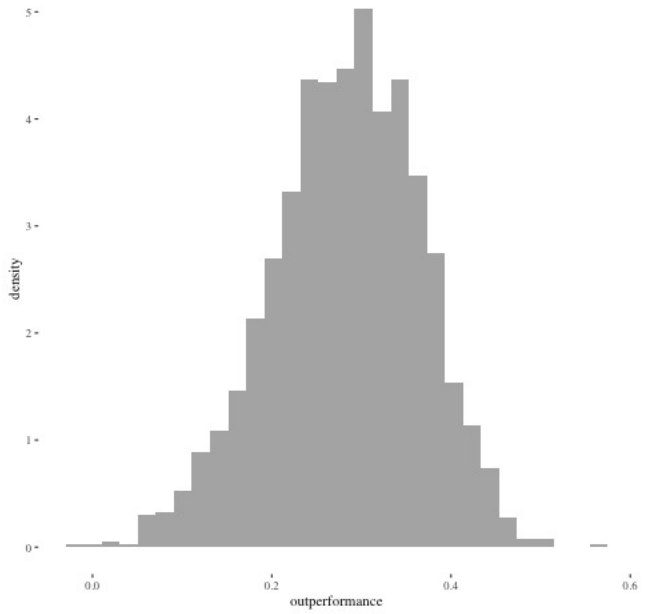

Figure 2.11. Bootstrapped estimates of returns to selectivity. This graph quantifies the distribution of performance gains derived from a bootstrapping procedure. akes the paired difference between the two distributions in (a) and therefore graphs the distribution of the benefit of being selective (up to 50%) relative to the chosen portfolio of VC investors (selectivity = 0%).

2.6 Unpacking VC decisions

In this section I explore the underlying decision model of VCs in order to understand why they invest in predictably bad investments.

2.6.1 Cream of the crop vs bottom of the barrel

In Section 2.5.1 I construct an algorithm to predict whether a startup will have a successful exit. In this section I will train algorithms for a different outcome: will the startup receive an early-stage investment. In particular, I will build an early-decision algorithm on the bottom 30% of firms according to the late-exit algorithm. Then I will build a second algorithm to predict early investments on the top 30% of firms according to the late-exit algorithm. Figure 2.12 plots the coefficients of each of the models and compares them to each other. Despite the fact that none of the coefficients carry causal interpretations, there are several

suggestive takeaways. First, nearly all points lie directly on the axes, suggesting that firms use totally different criteria when selecting good and bad investments. Second, the word tokens along the vertical axis appear to differ systematically from those along the horizontal axis—the model trained on the worst firms appears to prioritize founder details, whereas the model trained on the best firms appears to prioritize product details. When making good investments, investors appear to bet on the horse, but when making bad investments they appear to be betting on the jockey.

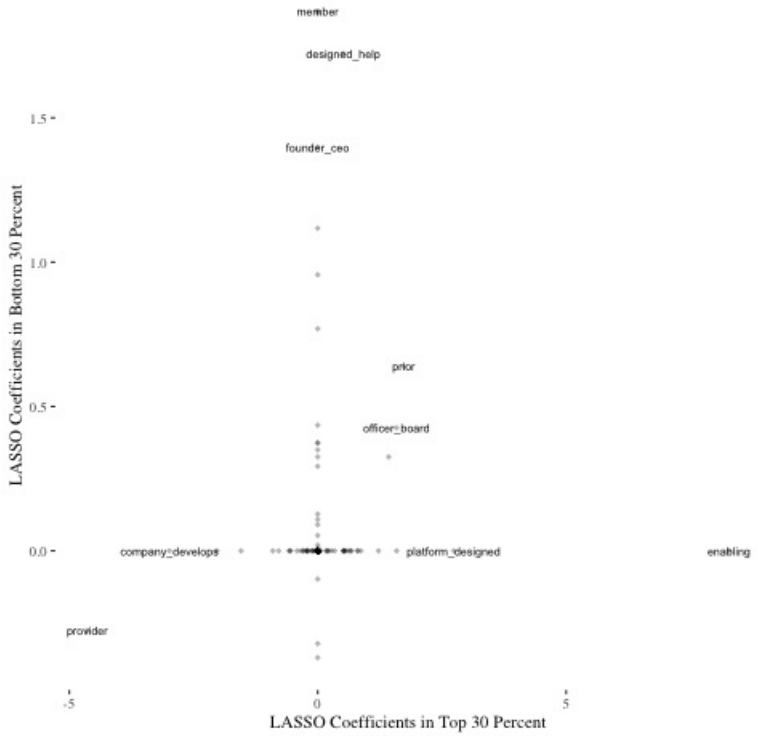

Figure 2.12. What Predicts Investments in Good vs Bad Investments? I build two LASSO models with the same feature set **X** to predict which companies will receive an early-stage investment: once on companies in top 30 percent of algorithmic success predictions (the wheat) distribution and once on companies in the bottom 30 percent of the algorithmic success predictions (the chaff). I plot the results of the two regressions where each point represents a variable, the x axis is the coefficient from the first regression and the y axis is the coefficient from the second. Nearly all points cluster around one of the axes though not at the origin, indicating the importance of features in one set is orthogonal to features important in the other. Investors have totally different criteria for the different quality firms. For example, "founder_serves" is a large, positive predictor for both sets. On the other hand "founded" is a strong negative predictor among the cream but it has no predictive power among the chaff.

2.6.2 A formal test of overweighting

Are investors relying too much on the jockey? Following Mullainathan and Obermeyer (2022), I test for overweighting of a founder cue by building an algorithm to predict late outcome strictly using the founder information. Then we can ask whether two companies that are equally likely to succeed according to X are differentially likely to receive investment based on $X_{education}$. Since X is a proper superset of $X_{education}$, the null corresponds to no overweighting of the founder information. Table 2.10 strongly rejects that null and suggests that investors overweight education in general, but especially for lower quality firms.

2.6.3 Are investors making mistakes?

I have shown evidence that indicates investors are systematically underperforming and systematically over-relying on certain signals of quality. The findings are consistent with misprediction and attentional biases and that is my preferred interpretation. However, other mechanisms could be at play and that would make it inappropriate to label these choices mistakes. For example, the incentive structure may not be well-aligned between principal and investor, and/or investors could have a rich set of preferences that includes reasons to tradeoff financial returns in favor of say interpersonal relationships, administrative ease, or simple effort reduction. These 'omitted payoffs' are known to complicate the interpretation of humans deviating from what an algorithm might recommend. For example, investors may get direct, non-monetary utility from investing in some kinds of companies or founders, or they may simply dislike the psychic costs of discerning between the best and worst companies they are likely to invest in. Further, orthogonal to their own preferences, investors may face external constraints from the capital providers. This could create a situation in which investors invest in the companies that are easiest to defend or justify.

2.6.4 The source and persistence of biases

I further develop my view on the underlying mechanisms by briefly exploring several plausible economic and institutional features in this setting that allow biases to persist. I argue here that the VC space represents a perfect storm of reasons why investors are likely unaware of their forgone returns and why they may not do much about them even if they did know.

Conventional wisdom

Business icons from Ray Kroc to Jeff Bezos have captured our attention and lend credence to the mythology surrounding founders and their startups. Hit TV shows such as "Shark Tank" and industry norms around "Pitch Day" events suggest an ethos built around founder charisma and personal persuasion. In his best-selling book "Good to Great", Collins (2009) asserts "First Who, Then What" and the idea has been absorbed into the zeitgeist. The ubiquity of this perspective is a true success story of "model persuasion" (Schwartzstein and Sunderam 2021). Investors seem convinced that the founder-first model of the world is the correct one. This likely facilitates investors neglecting to notice features that are predictive and a feedback loop of never noticing or learning persists, consistent with the model and evidence presented in Hanna, Mullainathan, and Schwartzstein (2014).

Performance is relative

Investors select an investment strategy, observe returns, and ask the question (in some form)—did we do well? The answer, of course, depends on the endogenous definition of success. What do investors compare their performance to? There are several intuitive candidates: an internal hurdle rate, last year's performance, investor's expectations, alternative investments, or one's competitors. This paper takes a different approach from all of these candidates. Instead of comparing the ex post returns of the chosen strategy to other ex post returns, I compare a set of available strategies using ex ante measures of quality. This choice

leads to a difference in sign as to whether investors are "outperforming". How investors choose the relevant counterfactual or reference group for performance is understudied in the literature and is ripe for future research. I argue that who and what investors compare themselves to is likely a significant influence on portfolio choice, especially if investors are "satisficers" and not maximizers (Simon 1955).

Feedback

Startup investing is a "wicked environment" (Hogarth, Lejarraga, and Soyer 2015). There is enormous selection in the distribution of outcomes we observe in the world, most success happens over long time horizons filled with factors and events that may seem relevant, and the underlying distribution of founders and companies changes over time. That an algorithm can find predictive signal suggests that a disciplined statistical process can find useful information in the large space of X, but the confounds of the environment may make those insights opaque to human investors.

Why don't investors build algorithms?

Despite their potential, we don't see investors universally adopting algorithms for investment decisions. There are at least two factors that could plausibly contribute to this lack of adoption: institutional constraints and behavioral preferences. For example, company inertia and preferences of limited partners could limit profit-improving technology adoption. Alternatively, investors may be overconfident in their abilities (Moore and Healy 2008), they may intrinsically value agency, or have an inherent algorithm aversion (Dietvorst, Simmons, and Massey 2015).

2.7 Conclusion

This paper offers evidence that sophisticated investors in high-stakes markets can sustainably make systematically poor choices. I offer one strategy to limit such bad decisions: have a human make a set of recommendations to an algorithm which then culls the options that are not likely to reach a benchmark threshold. Future research should explore interventions to reduce the misprediction of high-return opportunities and whether investors, their principals, or startup founders are the source of the suggested overemphasis on human capital.

Tables

Table 2.1

Summary statistics: Firm Characteristics by Early Stage

Sample statistics. Accelerator recipients define the sample. "Early VC" indicates the subsample of firms that received an early-stage investment after the accelerator; "No early VC" firms did not receive an investment. All fields correspond to data available when a given startup is accepted into an accelerator.

		No early VC		Early VC	
		Mean	Std. Dev.	Mean	Std. Dev.
Has late stage exit?		0.0	0.0	0.1	0.3
Number of funding deals		1.4	0.9	1.6	1.0
Accelerator funding amount ($M)		0.1	0.2	0.1	0.3
Accelerator stake (%)		16.4	16.6	13.6	13.8
Accelerator equity investment ($M)		0.1	0.2	0.1	0.3
Revenue ($M)		8.3	57.3	166.8	2310.5
EBITDA ($M)		-0.5	8.7	-0.5	8.2
		N	Pct.	N	Pct.
Has CEO bio?	FALSE	745	7.0	70	1.3
	TRUE	9869	93.0	5370	98.7
Female founder?	FALSE	9807	92.4	4880	89.7
	TRUE	807	7.6	560	10.3
Has CEO education?	FALSE	3817	36.0	868	16.0
	TRUE	6797	64.0	4572	84.0
CEO has MBA?	FALSE	9543	89.9	4591	84.4
	TRUE	1071	10.1	849	15.6
CEO has M7 MBA?	FALSE	10046	94.6	4866	89.4
	TRUE	568	5.4	574	10.6
CEO has PhD?	FALSE	9742	91.8	4928	90.6
	TRUE	872	8.2	512	9.4
Known industry code	FALSE	4	0.0	0	0.0
	TRUE	10610	100.0	5440	100.0
Has company description	FALSE	2	0.0	0	0.0
	TRUE	10612	100.0	5440	100.0

Table 2.2

Summary statistics: Firm Characteristics by Late Stage

Sample statistics. Recipients of early-stage funding define the subsample for this table. "Late-stage exit" indicates the subsample of firms that achieved a late-stage exit; "No late-stage exit" firms did not achieve an exit.

	No late-stage exit		Late-stage exit	
	Mean	Std. Dev.	Mean	Std. Dev.
Raised A or B round after accelerator?	0.1	0.3	0.4	0.5
Early-stage invested equity ($M)	0.5	2.3	3.9	6.7
Early-stage invested capital ($M)	2.1	4.5	3.9	6.7
Has IPO exit?	0.0	0.0	0.0	0.1
Has merger/acquisition exit?	0.0	0.0	0.8	0.4
Post-money valuation at exit			679.5	4182.5
Shut-down or bankruptcy?	0.3	0.5	0.0	0.2

Table 2.3

Top Accelerators

This table documents the 10 most successful accelerators in the data. Success is defined as the total five-year market value of all startups that have completed the accelerator's programming ("Alum Mkt. Value"). "% of Total" captures the percentage of all market valuations in my data that are attributable to a given accelerator. " of investments" captures the number of startups that complete a given accelerator and then raise an early-stage round within two years.

	InvestorName	Alum Mkt. Value ($M)	% of Total	# of investments
1	RocketSpace	51,000	0.53	3
2	Y Combinator	22,407	0.24	437
3	Plug & Play Tech Center	3,959	0.04	134
4	JLABS	3,686	0.04	38
5	MasterCard Start Path	1,700	0.02	2
6	Impact USA	1,400	0.01	4
7	Wharton Venture Initiation	1,200	0.01	13
8	StartX (US)	1,151	0.01	64
9	Agoranov	1,062	0.01	33
10	Techstars	771	0.01	280

Table 2.4

Top Investors

This table documents the early-stage investors associated with the most valuable firms. "Mkt.Value" captures the total five-year market value of startups for which a given investor served as the lead investor in the startups first funding round after the accelerator. "% of Total Mkt. Value" captures the fraction of all valuations in my data that are associated with an early stage investment by a given investor. "Invested" and "% of Invst." capture the total funding by an investor that I observe and the fraction that a given investor's funding constitutes of the total funding in my data, respectively.

	Investor Name	Invested ($M)	% of Invst.	Mkt.Value ($M)	% of Total Mkt. Value	Count
1	Benchmark	14.8	0.00	51,000	0.52	2
2	Sequoia Capital	67.9	0.01	5,564	0.06	6
3	Accel	200.6	0.03	5,348.48	0.05	11
4	Index Ventures	71.0	0.012	4,030.06	0.04	5
5	Andreessen Horowitz	139.2	0.02	3,720	0.04	18

Table 2.5

Top Accelerators

This table documents the 10 most successful accelerators in the data. Success is defined as the total five-year market value of all startups that have completed the accelerator's programming ("Alum Mkt. Value"). "% of Total" captures the percentage of all market valuations in my data that are attributable to a given accelerator. " of investments" captures the number of startups that complete a given accelerator and then raise an early-stage round within two years.

	InvestorName	Alum Mkt. Value ($M)	% of Total	# of investments
1	RocketSpace	51,000	0.53	3
2	Y Combinator	22,407	0.24	437
3	Plug & Play Tech Center	3,959	0.04	134
4	JLABS	3,686	0.04	38
5	MasterCard Start Path	1,700	0.02	2
6	Impact USA	1,400	0.01	4
7	Wharton Venture Initiation	1,200	0.01	13
8	StartX (US)	1,151	0.01	64
9	Agoranov	1,062	0.01	33
10	Techstars	771	0.01	280

Table 2.6

Top Markets

This table documents the representation of markets among different sets of companies. The first column of percentages shows the distribution of markets among all accelerator participants. The second column shows the distribution of markets among those companies receiving early stage investments. The third column shows the distribution of markets among the companies that achieved an exit.

	Market	#	% of All	% of Early	% of Late
1	Business/Productivity Software	2,038	0.127	0.186	0.167
2	Application Software	1,362	0.085	0.059	0.145
3	Media and Information Services (B2B)	834	0.052	0.061	0.056
4	Social/Platform Software	823	0.051	0.033	0.044
5	Information Services (B2C)	637	0.040	0.040	0.034
6	Financial Software	609	0.038	0.059	0.062
7	Electronics (B2C)	446	0.028	0.034	0.020
8	Educational Software	405	0.025	0.027	0.014
9	Other Services (B2C Non-Financial)	360	0.022	0.017	0.014
10	Entertainment Software	303	0.019	0.016	0.018

Table 2.7

ML-OLS Projection, Levels

This table reports regression results to approximate the relationship between the algorithm's prediction and simple variables previously explored in the literature. The algorithm's success prediction is the predicted probability of success $p \in [0, 1]$.

	Dependent variable:				
	Algorithm predicted success				
	(1)	(2)	(3)	(4)	(5)
(Pre-accel. funding)^2	-0.00002^{***}				-0.00002^{***}
	(0.00000)				(0.00000)
MBA CEO		0.030^{***}			0.027^{***}
		(0.001)			(0.001)
M7 MBA CEO			0.020^{***}		0.009^{***}
			(0.002)		(0.002)
Female founder				0.007^{***}	0.005^{***}
				(0.001)	(0.001)
Pre-accel. funding	0.003^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.002^{***}
	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
Constant	0.047^{***}	0.044^{***}	0.046^{***}	0.047^{***}	0.042^{***}
	(0.0004)	(0.0004)	(0.0005)	(0.0005)	(0.0005)
Observations	4,578	4,578	4,578	4,578	4,578
R^2	0.041	0.141	0.057	0.027	0.166
Adjusted R^2	0.041	0.141	0.057	0.027	0.166

*p<0.1; **p<0.05; ***p<0.01

Table 2.8

ML-OLS Projection, Percentile

This table reports regression results to approximate the relationship between the algorithm's prediction and simple variables previously explored in the literature. The subsample used for this analysis excludes startups that did not receive early-stage investments. The dependent variable for Columns (6)-(10) is the algorithm's predicted probability of success p. The dependent variable for (1)-(5) is the percentile of p (predicted probability of success) within the set of firms that received early-stage investments.

	Dependent variable:				
	ML Predicted Success Percentile				
	(1)	(2)	(3)	(4)	(5)
(Prior funding)^2	−0.001***				−0.001***
	(0.0001)				(0.0001)
MBA CEO		0.211***			0.189***
		(0.017)			(0.018)
M7 MBA CEO			0.137***		0.058***
			(0.021)		(0.021)
Female CEO				0.028	0.021
				(0.022)	(0.021)
Prior funding	0.048***	0.018***	0.018***	0.018***	0.047***
	(0.005)	(0.002)	(0.002)	(0.002)	(0.004)
Constant	0.555***	0.541***	0.558***	0.570***	0.519***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	1,535	1,535	1,535	1,535	1,535
R^2	0.069	0.121	0.061	0.035	0.157
Adjusted R^2	0.067	0.120	0.060	0.034	0.154

*p <0.1; **p <0.05; ***p <0.01

Table 2.8 (continued)
ML-OLS Projection, Percentile

This table reports regression results to approximate the relationship between the algorithm's prediction and simple variables previously explored in the literature. The subsample used for this analysis excludes startups that did not receive early-stage investments. The dependent variable for Columns (6)-(10) is the algorithm's predicted probability of success p. The dependent variable for (1)-(5) is the percentile of p (predicted probability of success) within the set of firms that received early-stage investments.

Dependent Variable:					
	ML Predicted Success Probability				
	(6)	(7)	(8)	(9)	(10)
(Prior funding)^2	−0.0001***				−0.0001***
	(0.00002)				(0.00002)
MBA CEO		0.028***			0.028***
		(0.002)			(0.002)
M7 MBA CEO			0.012***		0.001
			(0.003)		(0.003)
Female CEO				0.001	−0.0001
				(0.003)	(0.003)
Prior funding	0.005***	0.002***	0.002***	0.002***	0.005***
	(0.001)	(0.0003)	(0.0003)	(0.0003)	(0.001)
Constant	0.051***	0.049***	0.052***	0.053***	0.047***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1,535	1,535	1,535	1,535	1,535
R^2	0.048	0.120	0.035	0.021	0.144
Adjusted R^2	0.047	0.119	0.033	0.019	0.141

*p <0.1; **p <0.05; ***p <0.01

Table 2.9

Bootstrapped Confidence Intervals for Quadratic Relationship

This table reports the bootstrapped confidence intervals for the coefficient on the quadratic term in the relationship between ex ante selectivity and returns (MOIC). Negative signs indicate a bill shaped relationship. For each confidence level, I calculate the bootstrapped interval by three different procedures. * indicates the interval does not include 0.

Confidence Level	Normal	Percentile	BCa
90%	(-5.405, 0.670)	(-5.708, 0.335)	(-6.939, -0.170)*
95%	(-5.987, 1.252)	(-6.459, 0.726)	(-8.019, 0.171)

Table 2.10

Human capital overweighting

This table reports regression results from correlating several algorithmic success predictions to a binary dependent variable indicating if the startup received an early-stage investment (=1) or not (=0). The "Full model prediction" is the algorithm's predicted success when trained on all available variables. The "Edu model prediction" is the algorithm's predicted success when trained only on founder's education information. The non-linear risk controls add fixed effects for the quintile of ML-predicted success.

	Dependent variable:				
	Received early stage investment?				
	(1)	(2)	(3)	(4)	(5)
Full model prediction	3.430***	2.783***	7.107***	0.631*	2.371***
	(0.175)	(0.219)	(0.476)	(0.334)	(0.785)
Edu model prediction		1.357***	4.990***	0.582**	1.820***
		(0.275)	(0.449)	(0.281)	(0.578)
Full*Edu model prediction			−62.335***		−19.079**
			(6.097)		(7.794)
Constant	0.213***	0.165***	−0.050**	0.146***	0.075**
	(0.008)	(0.013)	(0.025)	(0.016)	(0.033)
Non-linear Risk Control?	No	No	No	Yes	Yes
Observations	10,452	10,452	10,452	10,452	10,452
Adjusted R^2	0.035	0.037	0.047	0.056	0.057

*p<0.1; **p<0.05; ***p<0.01

CHAPTER 3

ALGORITHMIC CURATION CREATES BIAS: THEORY, EXPERIMENT, AND EVIDENCE FROM FACEBOOK

3.1 Introduction

Most discussions of algorithmic bias omit, or leave underdeveloped, a culprit: people. Of course, the role of human agency is widely acknowledged. After all, algorithms are designed by and trained on data generated by humans. Still, much of the empirical and theoretical research focuses on the intricate mathematics of algorithm construction or the statistical vagaries of datasets. The rich underlying psychology of people is usually abstracted away. In this paper, we draw out the consequences for algorithmic bias of a single robust fact about human psychology, using both lab experiments and an empirical audit of two very large algorithms that affect billions of people.

The psychology we focus on is a foundational one for modern behavioral science: we choose differently when thinking fast than when thinking slow. In broad strokes, when choosing automatically, our choices may not reflect our true, reflective preferences. When it comes to discrimination, this insight has - again, in broad strokes - an implication: we are more prejudiced when we choose quickly than when we choose slowly. Many of the forces that create discrimination operate quickly - stereotypes, gut responses - which are then muted on more careful reflection. Throughout this paper, takes deliberate preferences as the benchmark and calls outcomes that are more prejudiced than those as bias.[1]

Even stated so abstractly, the recognition of automaticity proves useful, by focusing attention on an overlooked aspect of the behavioral data used to train many machine learning algorithms. We often build algorithms that take data on our past choices, infer from them

1. Notice, even those deliberate choices can contain prejudice. Our focus is on the *additional* gaps that arises from behaving automatically. In addition, notice that we are abstracting here from the possibility that stated preferences are not our actual preferences but merely the ones we state to placate a surveyor.

our preferences and then proceed to make recommendations, curate choice sets or even automatically choose for us. Yet the inferences those algorithms will draw about what we want will depend on how we made those choices - deliberately or automatically. This produces a clear implication: the extent of algorithmic bias depends on the automaticity of choice in the training data. Algorithms trained on more automatic choices will infer greater prejudice on our parts (than we actually wish to be on reflection) and proceed to act on it. We test this implication for a particular kind of prejudice: our tendency to favor people like us, favoring "own group" members over "out group" members.

Our theory predicts that human bias—and hence algorithmic bias—should be most pronounced in contexts in which decision-making is relatively more automatic and less deliberative. Importantly, note that this distinction is almost a philosophical one without access to the data on the variation in context. Without seeing inconsistent choices across fast versus slow contexts, we'd have no way of knowing if a user simply gets greater utility from own-group posts or if they were instead exhibiting implicit bias.

Our first empirical test of that hypothesis comes from a "natural experiment" embedded within one of the world's largest social media platforms: Facebook. We first carry out a survey of Facebook users and ask them about their decision-making as it relates to two algorithms on the Facebook platform, *News feed*, which ranks the posts of a user's friends, and *People You May Know* (PYMK), which ranks potential new friends. We show using a variety of metrics that users report more automatic behavior when scrolling through posts (the data used to train NewsFeed) than when scrolling through potential friends (PYMK's data). We then perform empirical audits of US facebook users and, consistent with our theory, find significant outgroup bias in NewsFeed. Even holding constnat the user's *explicit* preferences, own-group content is up to 25% more likely than out-group content to be ranked by Newsfeed in the top five posts shown to users. And, also consistent with our theory, we find no detectable disparity for the PYMK algorithm.

To further test the general validity of these results, we conduct an additional audit in India. In this context, own-groups and out-groups are defined very differently, no longer by race but instead by religion: Hindus versus Muslims. Despite the change in geography and a focus on religion, we again find the same pattern of results. Newsfeed rankings are biased against Hindu posts by friends of Muslim users, and biased against Muslim posts by friends of Hindu users. And, again, we find no detectable evidence of bias with the PYMK algorithm.

Because other things differ between Newsfeed and PYMK besides automaticity of the choice context, we carried out a third fully-controlled test in the lab. We develop a task where subjects select movies recommended to them by (fictional) strangers, who are randomly assigned an indicator of own- versus out-group status (name) following Bertrand and Mullainathan (2004). Subjects on average have a preference for movies recommended by own-group members, especially when choosing in the randomly assigned "rushed" condition. We also show that an algorithm built using subject choices as its training dataset shows more out-group bias in rankings when trained using data from the rushed condition than the non-rushed condition. And then in a final lab experiment we show using this algorithm to curate (rank-order) people's choice sets leads to more biased choices compared to when people choose from randomly-ranked recommendations.

The field versus lab studies have offsetting strengths and weaknesses. The control possible in the lab setting lets us pinpoint automaticity as a key mechanism for algorithmic bias, but leaves open the question of external validity. The Facebook audits have more room for confounding, but suggest the immense practical relevance of our theory given Facebook has 2.9 billion users per month and is a source of social capital of the sort that affects a wide range of life outcomes.[2] Taken together these studies paint a compelling, but concerning,

2. Putnam 2000 distinguishes between "bonding" social capital, or "strong ties" with friends and families that provide emotional and other supports, and "bridging" social capital, or "weak ties" (Granovetter, 1973), that provide people with valuable new information and perspectives. The advent of social media has added a third category to this list, "maintained" social capital, or the ability to perpetuate ties to people with

picture.

Our results argue for much greater attention to the question of what behaviors go into the training data used to construct algorithms. In many decision contexts several different kinds of behaviors are observed, such as whether to hover for an extra millisecond before scrolling by, or to "like" a post, or even to write a response. These behaviors can happen at different levels of mental deliberation, and so—our results suggest—can vary quite a bit as to their alignment with what users actually want. The implication is that the design of human-facing algorithms must be as attentive to the underlying psychology of the human users as to the machine learning engineering.

3.2 Conceptual Framework

We build a simple formal framework to state more precisely our core assumptions and the implications we draw out. We focus on the problem of curation. Users face a large number of potential options (a set S): posts, tweets, products movies, even job applications. Users must sift through this large set of options to find the ones they like. The fundamental challenge of curation is to help the user better sift. In our model, we will assume the reader goes through the items one by one; that is, the set is ranked and the user starts with the highest ranked item and proceeds downward. Each piece of content has features x_s (for example, length, topic, etc.), and a binary feature g_s for whether it was produced by an out-group member (1) or in-group member (0).

Engaging with s produces a real valued utility $u(s)$, which is what we mean by user

whom one has lost face-to-face contact (Ellison, Steinfield, and Lampe, 2007). Existing research suggests that use of Facebook at all relative to no use, relatively more intensive use of Facebook, and investments in time on "Facebook Relationship Maintenance Behavior" are all associated with increased social capital, particularly the weak ties associated with bridging social capital (Antheunis, Vanden Abeele, and Kanters, 2015). Previous research has found that weak ties are positively related to important outcomes like creativity (Baer, 2010), employment status and income (Tassier, 2006), risk of crime involvement (Patacchini and Zenou, 2008), health (Kawachi, Berkman et al., 2000), and subjective well-being (Sandstrom and Dunn, 2014). Previous studies also suggest that intergroup contact over social media, including on Facebook specifically, may reduce prejudice (Alvídrez et al. 2015; **?**).

preferences. If utility were the only determinant of user choices and out-group posts generated less utility for users ($E[u(s)|g_s = 1] < E[u(s)|g_s = 0]$), then an algorithm trained on user choices would lead to ranking of content $r^u(s, S)$ (with highest-utility content ranked 1 and so on) that would simply reflect user preferences.

The fundamental challenge arises from the fact that the data typically used to train algorithms includes not information about our preferences, but instead information about our choices—that is, not $u(s)$ but rather whether we engage with (or "click") a piece of content, $c(s)$. A large body of research from psychology demonstrates that our choices can often deviate from our preferences ("intention-action gaps"), especially in decision settings that are "fast" ($f = 1$) where our cognition is relatively more automatic than in settings where our choices are relatively more deliberate or slow ($f = 0$).[3] Kleinberg, Mullainathan, and Raghavan (2022) show how this phenomenon can affect our choices about consumer goods. We extend that framework here to consider the implications of user bias, b_f. Let the decision to engage with or click content obey:

$$Pr(c(s) = 1) = \frac{e^{u(s)(1-b_f g_s)}}{1 + e^{e^{u(s)(1-b_f g_s)}}}$$

The psychology research implies $b_1 > b_0 \geq 0$ so that fast contexts ($f = 1$) have greater bias than slow contexts ($f = 0$).[4]

Since utility is unobserved the algorithm ranks by choices instead, which we call $r^{c,f}(s, S)$. This is sorted by $c(s, f)$ with the most frequently-clicked item ranked $r(s) = 1$ and so on. We define the *disparity* in a ranking rule r as the expected difference in ranking for in-group versus out-group posts, $\Delta(r^u) = E[r(s)|g_s = 0] - E[r(s)|g_s = 1]$. It is easy to see that ranking by user behavior or engagement rather utility increases favoritism for in-group

3. See, for example, Kahneman (2011) and Kahneman, Sibony, and Sunstein (2021).

4. See for example Payne, Lambert, and Jacoby (2002), Lueke and Gibson (2015) and the studies reviewed there.

content (because of intention-action gaps), and that these disparities are even larger when engagement is measured in fast contexts.

$$\Delta(r^{u}) < \Delta(r^{c,f=0}) < \Delta(r^{c,f=1})$$

Algorithmic ranking increases disparity above and beyond the problem of using engagement as a proxy for preference. The algorithm ranks on *predicted* engagement, which is formed from a data set of many pairs of the type (c, x, g) in order to predict engagement for new posts for which we have just (x, g) available. Let us generously assume the algorithm makes the best possible prediction given an infinite number of data points, so for any post the algorithm perfectly predicts $E[c(s, f)|x_s, g_s, f]$. So the algorithmic ranking $r^{a,f}(s, S)$ results from sorting the set of posts by $E[c(s, f)|x_s, g_s, f]$. Algorithms have more bias when trained on data from fast than slow contexts:

$$\Delta(r^{a,f=0}) < \Delta(r^{a,f=1})$$

But they do not simply replicate the bias in engagement. They *magnify* the bias: algorithmic ranking adds an *additional* disparity. The algorithm does not know the actual click rate for every post $c(s)$. Instead it must use the *expected* click rate; that is, the average click rate of similar posts, those with the same (x, g). As a result, it is in effect stereotyping the posts: outgroup posts are lumped with other posts. If there is bias against some outgroup posts, all outgroup posts can wind up penalized. Worse still, because the algorithm pools data across users, it propagates implicit biases across people. If any users are biased then:

$$\Delta(r^{u}) < \Delta(r^{c,f}) < \Delta(r^{a,f})$$

.

The remainder of our paper tries to empirically test this hypothesis.

3.3 US Facebook Audit

To empirically test our theory we carried out audit studies of two Facebook algorithms, Newsfeed and People You May Know (PYMK), in two countries, the US and India. We show that users report spending more time deliberating before making choices on PYMK compared to Facebook, which makes sense given the relatively higher stakes of choosing to friend someone on Facebook relative to whether to click on a post on Newsfeed. Our model predicts that we should see more pronounced algorithmic bias with Newsfeed than with PYMK. That is exactly what we find in both countries. Moreover the magnitude of Newsfeed's bias is substantial.

3.3.1 Automaticity Across Facebook Algorithms

A key prediction of our model is the idea that behavioral biases - and hence algorithmic biases - are most pronounced when behavior is not guided by deliberate thought. We hypothesized that Facebook would provide a sort of "natural experiment" to test that hypothesis, since the Newsfeed algorithm curates a large number of low-stakes choices for users (how to rank-order posts from friends) while the PYMK algorithm curates a smaller number of higher-stakes choices (whether to "friend" another user on the platform or not).

To test our hypothesis we first carried out a survey in the US (on the Prolific platform) to measure the amount of cognitive effort, deliberation, and time spent making choices to interact with posts on Newsfeed versus add a friend from PYMK (N=300). We draw on existing measures in the literature about, for instance, how well the subject could explain their choices, how much "mental effort" they say they put into the behavior, whether the decisions are based on "gut feelings" or careful consideration, and how much time they usually spend (in seconds) making the decision (Bargh, 1994).

Figure 3.1 shows that for each of the nine measures of how deliberate people's interactions are with Facebook's algorithms, users report higher levels of deliberation with PYMK than Newsfeed. The advantage in favor of PYMK in terms of how deliberate the behavior is range from 0.05 standard deviations (for inattention) to 0.53 standard deviations (for carefulness). A simple composite index of the standardized measures suggest a "deliberation advantage" of PYMK over Newsfeed of 0.22 standard deviations. When we do a principal components analysis across the measures and compare the first principal component across algorithms, the PYMK advantage over Newsfeed in terms of being more deliberate equals 0.4 standard deviations. And Figure 3.2 shows the CDF for responses to the one continuous measure, time it takes the study subject to make a decision (in seconds), which again shows the same pattern.

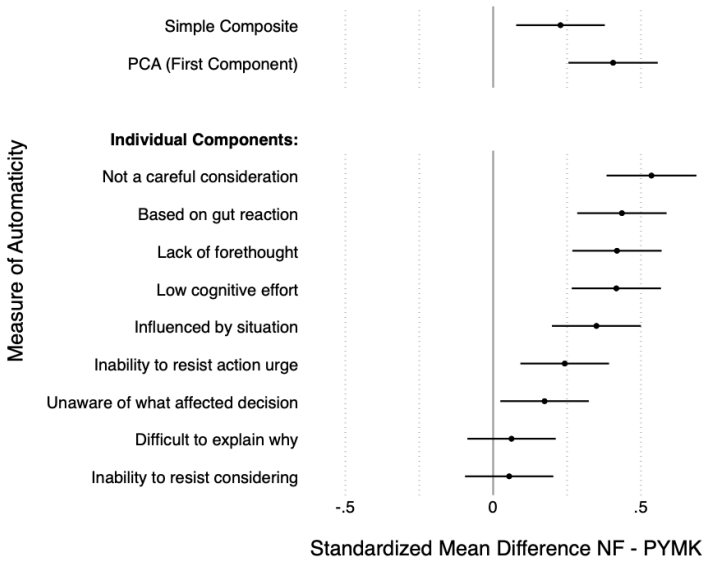

Figure 3.1. Standardized Effect Sizes. In a survey, we collected 10 measures of deliberateness in making decisions to engage in content on the Newsfeed and the People You May Know recommendations.The standardized effect size is the mean response for the newsfeed and the mean response for PYMK divided by the pooled standard deviation. These are shown for NF − PYMK for the various measures of automaticity. Higher values indicate more automatic for all questions except time where more seconds indicates more such that values > 0 imply more automatic decision making in Newsfeed versus PYMK. The first two measures are a simple composite average of the underlying individual components and the first principal component of those underyling individual measures.

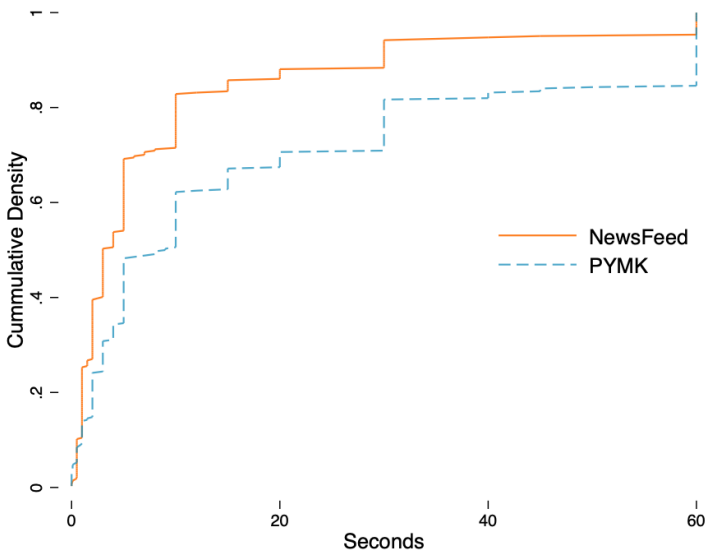

Figure 3.2. Speed (time to decide in seconds) CDF. In a survey, we collected 10 measures of deliberateness in making decisions to engage in content on the Newsfeed and the People You May Know recommendations. The graph shows the CDF of self-reported typical time to decide to take an action on deciding to engage in content on Newsfeed and PYMK (top-coded at 60 seconds). Ingroup is defined as same race.

3.3.2 US Facebook Audit Design

We advertised for study subjects who were Facebook users and were willing to participate in a Zoom-based interview. Subjects were first asked to complete a survey asking about their demographic characteristics and basic Facebook usage patterns. Subjects were then asked to log in to their Facebook account and share their screen. Data was collected in several waves, and different subjects had different data collected depending on the wave they participated in:

- For all subjects, enumerators captured the Newsfeed algorithmic ranking, and information about each of the first 60 posts in the user's Newsfeed (N=662).
- One subset of Newsfeed subjects then participated in a similar data collection about the first 60 friend recommendations from the PYMK algorithm (N=436)

- A different subset participated in data collection about their 10 most recent *interactions* with Newsfeed posts on Facebook (N=104)

- A third subset had no further data collection (N=122)

Which data collection a subject participated in was determined by the date and location of data collection. In our main analysis we will use the full samples available for each result—that is, we present results for Newsfeed for all observations for which we collected Newsfeed data, results for PYMK in the full sample for which we collected PYMK data, and results for interactions with Facebook posts in the full sample for which we collected interaction data.

For the US study we define own-group and out-group by whether the user and the Newsfeed poster (or potential friend being recommended) belong to the same race/ethnic group. So for example, if a white subject has a post from a white friend, a Black subject has a post from a Black friend, or an Asian subject has a post from an Asian friend, those would be coded as "own-group." To avoid priming study subjects about the topic of race, and because of time constraints on our data collection with each subject, we did not ask subjects themselves to report the race/ethnicity of posters on their Newsfeed. Instead we asked our enumerators, who could see the subject's Facebook account through screen sharing, to record their perception of the race/ethnicity of the posts from people (not companies) in each of the first 60 total posts and first 60 friend recommendations.

We asked subjects to self-report their own race and ethnicity using the seven-category system from the US Census, where subjects can check as many boxes as they like. We also asked the enumerators to record their perception of the race/ethnicity of the subject before the Facebook data collection began, which match subject self-reports 85% of the time.[5]

5. This is under a rather strict definition of match: enumerator and subject race choices had to match exactly. However, subjects often chose more than one race but enumerator's rarely did so. When we expand the match to be that the enumerator indicated the same race as at least one of the responses of the subject then the match rate is even higher.

We also collect direct measures of people's explicit utility or preferences:

- For Newsfeed, we asked subjects to report for each of the posts from people amongst the first 60 total posts they see: "There are more posts than Facebook can possibly show you. How would you rate this post on a scale from 1-7 where 1 means 'can skip' and 7 means 'definitely want to [see]"'.

- For those that participated in the PYMK data collection, for the first 60 friend recommendations on PYMK they are asked: "How familiar are you with this person on a scale from 1-7?" where 1 is not familiar and 7 is very familiar."

The enumerators also record ancillary information about the Newsfeed posts such as how long ago it was posted, whether the post is made by a person versus company, and whether the post was to a Facebook group. For PYMK recommendations, enumerators recorded additional information like how many mutual friends that each friend recommendation had with the subject.

As noted above some subjects were asked about not just preferences for posts, but also their recent behavior on Facebook (N= 104). Specifically we recorded the 10 most recent posts on Newsfeed that the user had some *interaction* with ("liking" or choosing another reaction or commenting), and exactly what action they took. Enumerators then also recorded the perceived race/ethnicity of the poster.[6]

We collected Newsfeed information from (N=662) subjects, PYMK information from (N=436) subjects, and recent Newsfeed interactions from (N=104) subjects. Our subjects are on average in the mid-20s (mean= 26.6, sd= 9.186, range=18 to 69). A large fraction of our sample check Facebook at least weekly. Compared to all US adults, our samples tend

6. The algorithm presumably has access to a wider range of behaviors than this, such as how long the user lingered on a post, whether the user clicked to expand on the post text or comments, whether the user watched a video and how much of the video was watched, etc. Given the constraints on our data collection, interacting with posts was the most feasible measure of actual user behavior on the network.

to have a higher proportion Asian, female, and people with a four-year (bachelor's) college degree.

Note that subjects were not made aware that this study was about race or own-group biases, and what data was exactly being recorded by the enumerator was unknown to the subject. For 50 subjects, at the end of the survey we asked "What do you think the purpose of this study is?" Not one mentioned race, gender, or other indications of own-group biases.[7]

3.3.3 US Facebook Audit Study Results

Our model suggests that the behavioral wedge (relative to preferences) in the direction of own-group bias should be larger for the Newsfeed algorithm (where decision-making is more rushed and automatic) than for PYMK (where decision-making seems to be more deliberate). This is indeed what we find in our audit study of Facebook users in the US.

Figures 3.3 and 3.4 shows these results for both Newsfeed and PYMK. We first normalize people's explicit preference ratings for Newsfeed and PYMK recommendations to account for the fact that different people use the Likert scale differently, and in particular we see systematic differences in Likert scale distributions for white and Black study subjects.[8] An own-group post in the bottom quartile of the user preference distribution has a *higher* Newsfeed ranking than an out-group post in the *next-highest* preference quartile. In a regression, own-group posts are ranked by Newsfeed 1.19 slots higher than outgroup posts even conditional on user preference, a difference that is statistically significant at the 5% level (with a standard error of 0.208). In contrast, we see no detectable differences in PYMK rankings conditional on user preferences. Our 95% confidence intervals let us rule out an

7. Similarly, enumerators were not told the purpose of the study, though they were of course aware they were collecting information on race.

8. Specifically, Black subjects tend to be more likely to use higher Likert preference rankings, indicating that they more prefer a post. Black subjects also see fewer posts by own-group friends on average (48% versus 51% for non-Black subjects. The normalized preferences re-scale scores relative to each subject's own average reported preference, and so take into account differences across subjects in the use of the Likert scale.

own-group effect on PYMK rankings any larger than 0.187 slots.

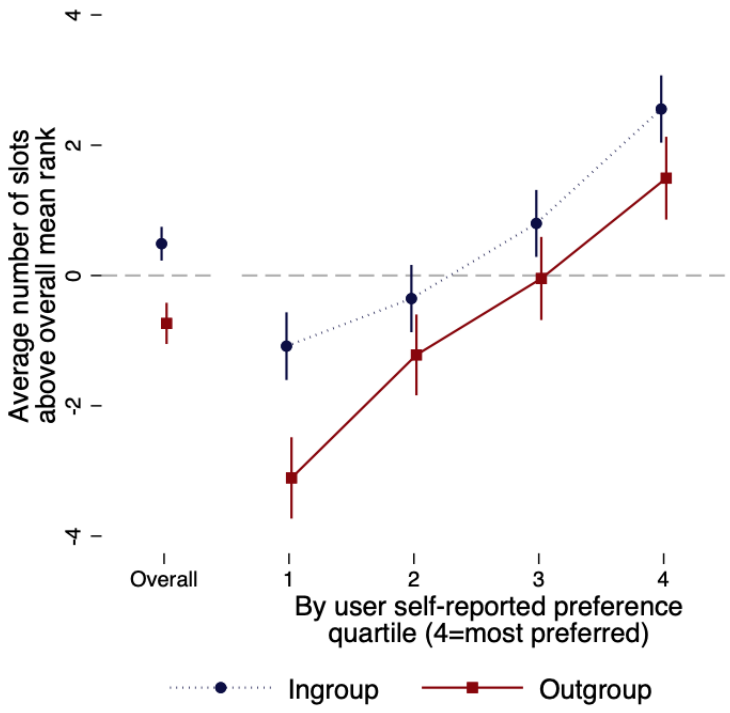

Figure 3.3. Relationship between Newsfeed Algorithmic Ranking and In-group Status Conditional on Subject Explicit Preference. We show the mean ranking of in-group and out-group posts above the overall mean, and then we show this by subject stated preference. The normalized subject explicit preference quartile is the across subject quartile of within subject z-scores for stated preference for a post with a suggested friend. Each subjects ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins. Ingroup is defined as same race.

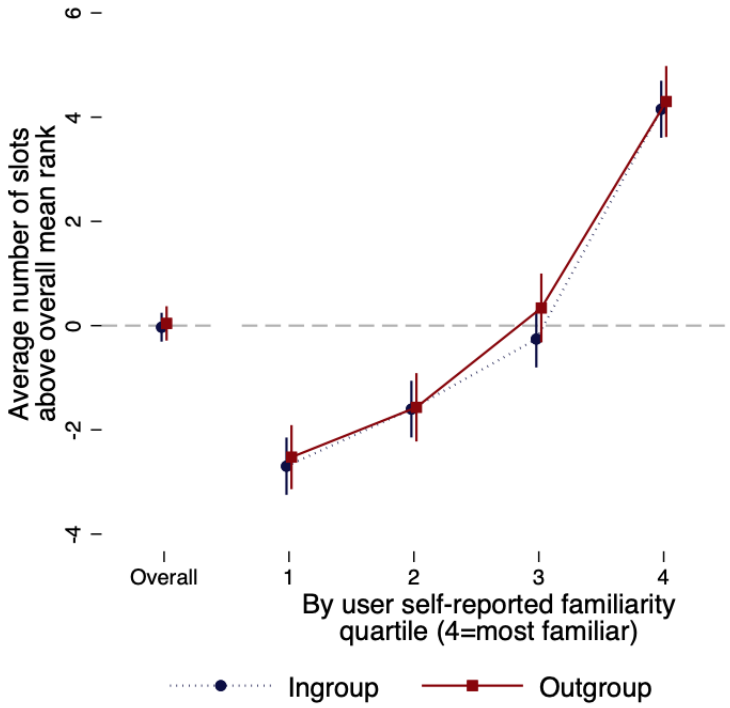

Figure 3.4. Relationship between PYMK Algorithmic Ranking and In-group Status Conditional on Subject Explicit Preference. We show the mean ranking of in-group and out-group recommendations above the overall mean, and then we show this by subject stated familiarity. The normalized subject explicit familiarity quartile is the across subject quartile of within subject z-scores for stated preference for a familiarity with a suggested friend. Each subjects ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins. Ingroup is defined as same race.

Our results do not seem to be an artifact of our particular estimation choices. We see similar results if we look at the chances a post winds up ranked in the top 5, 10 or 20 slots. One may worry this is about family members, if Facebook is aware of which existing friends are a family member and upranks those posts, and family members are highly likely to be "ingroup" then our results could reflect family status. In Wave 5 of our data collection we gathered information on how participants knew the poster, and in omitted supplemental work we use these data to interact own-group status and preference with a binary indicator for family. We see that even amongst non-family posts, Facebook upranks own-group posts

even conditional on user explicit preferences. Our results are also not simply due to how we normalize measures of explicit preferences; when we instead estimate the rank correlation between algorithmic ranking and explicit preferences,[9] the rank correlations with Newsfeed equal 0.11 for own-group posts and 0.07 for out-group posts, while for PYMK the correlations are very similar for own-group and out-group recommendations, equal to 0.11 and 0.12, respectively.

Interestingly, even though Newsfeed rankings seem to have an own-group bias, our measures of people's explicit preferences about Newsfeed content does not. Work completed in omitted supplemental materials shows that for normalized self-reported user preferences, the quartile rankings of own-group and out-group posts are very similar; none of the differences are statistically significant at the 5% level. In omitted supplemental work we also show a CDF of the normalized preferences people report for own-group and out-group content, which basically fall right on top of each other. We might worry about social desirability bias in subject responses, but subjects did not know the study was about race. Moreover, for subjects to intentionally misreport and hide any explicitly biased preferences would require them to keep a running tally of their reported preferences across all the top 60 Newsfeed posts separately by in-group and out-group.As noted above, for 50 participants we asked them what they thought the study was about, not one mentioned "race", "bias", "discrimination", "ingroup" or any similar phrases. The most straightforward explanation for these results is that algorithms that learn preferences from quick decisions can become biased even from users with no explicitly biased preferences.

Notice another implication, and testable prediction, of our model. The algorithm learns the biases of users, particularly in fast decision-making contents, and as a result up-ranks own-group content relative to out-group content holding constant the user's explicit preferences.

9. We estimate Kendall's tau-b correlation coefficient to deal with the fact that with a $1-7$ scale for explicit preferences we will have a large number of ties across the 60 posts we collect from each subject. Our thanks to Evan Rose for this suggestion.

In other words, the biased algorithm will compound the user's own bias (or bias of users like her) by showing the user a choice set that over-represents own-group items. So we would expect the magnitude of the bias in user's *behavior* should be even larger than the bias we see in the algorithm rankings, given that users are choosing from an already-biased set of options. In other words biased algorithms affect $P(choiceisseen|race)$, and biased humans affect $P(chosen|seen, race)$. The result is a "double penalty" against out-group content, especially for choices made in rushed decision-making contexts.

To explore this hypothesis we collected data on the last 10 *interactions* people had with Newsfeed posts: reactions and comments. This is not a perfect measure of user behavior with Newsfeed, since there are other behavioral dimensions that we cannot mention in our setting (like the time the user had spent looking at each post, which we cannot capture in our lab setting). We present results here just for the sub-sample of respondents for which we have this behavioral measure.

Figure 3.5 shows that the share of Newsfeed posts that the user has interacted with that are in-group is indeed larger than the share of posts that are in-group among those that the algorithm ranks towards the top of the user's Newsfeed. The confidence intervals here are somewhat large but the excess share of in-group posts among those the user interacts with (user behavior) is between 50% and 100% higher than the posts the algorithm puts in the user's top 5, 10 or 15 of the Newsfeed ranking. The algorithm is not merely reflecting our own behavioral biases back to us, it is amplifying those biases in our behavior.

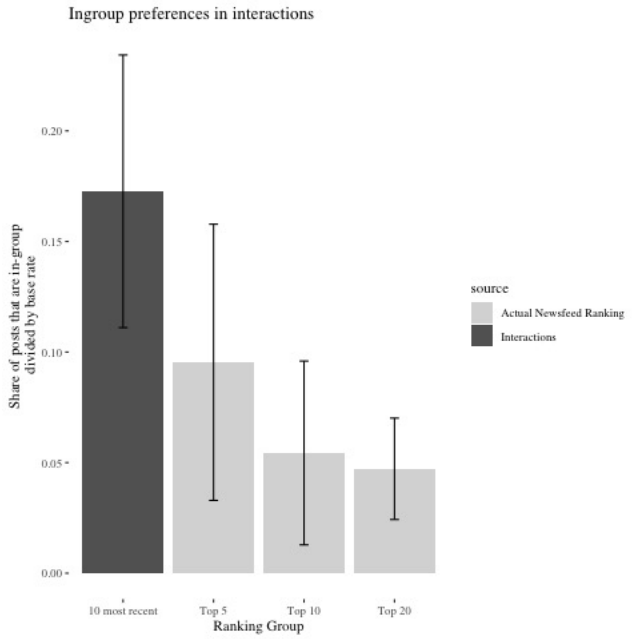

Figure 3.5. Share Recent Interactions In-Group and amongst Newsfeed Posts. For a subset of individuals, we collected information on the last 10 posts they had actually interacted with on Newsfeed (this does not necessarily have to be any of the posts currently on their Newsfeed). Recent interactions include the 10 most recent "likes", reactions, and comments. This figure shows the overrepresentation (above base rate) of ingroup posts in recent interactions (dark grey) and on the Top 5, 10, and 20 posts on the current observed Newsfeed (light grey). Ingroup is defined as same race.

3.4 India Facebook Audit

Part of what makes Facebook an interesting test-case for our hypothesis is its massive scope. Billions of people around the world regularly use Facebook and rely on its algorithms. So far we have presented evidence for the world's second-largest Facebook market, the US. But the US, with its particular cleavages by race and ethnicity, accounts for just 10% of all Facebook users world-wide.[10] Do our results hold more generally?

To answer this question we replicated our audit study in the world's single largest Facebook market: India. Rather than define own-group/out-group by race, which is signaled by user images in the US Facebook context, we define this now by religion, which is signaled

10. https://worldpopulationreview.com/country-rankings/facebook-users-by-country

by name in the Indian Facebook context.[11] Our results are qualitatively similar in India as in the US.

3.4.1 India Facebook Audit Study Design

Other than defining in-group/out-group status by religion, the study design of our India Facebook audit proceeded identically as in the US.

We recruited $N = 200$ study subjects via the Ashoka University Centre for Social and Behaviour Change (CSBC). Omitted supplemental work shows that our sample somewhat over-represents men (60.7% or our study sample) as well as Muslims (30.1% of our sample, compared to 14.2% of the general population). In omitted supplemental work we show the results we present below are qualitatively similar when we re-weight our analysis to make our weighted sample more nationally representative by gender and religion.

3.4.2 India Facebook Audit Study Results

In Figures 3.7 we show that, as in the US data, there is a sizable difference in Newsfeed post rankings for own-group content (defined by religion) relative to out-group content, even conditional on explicit user preferences. For example, own-group posts in the bottom quartile (least preferred) of user preferences have an average Newsfeed ranking that is not all that different from the average Newsfeed ranking for outgroup content in the third (next-to-most preferred) quartile.

11. Names are quite distinct in India for Hindus (roughly 80% of the country's population) versus Muslims (14% of the population). Like race in the US, religion is a fraught fault line in Indian society. Hundreds of thousands of people died when Muslim and Hindu populations were partitioned in 1946 into the countries of Pakistan and India, respectively. Discrimination and even violence on the basis of religion remains common in India to this day.

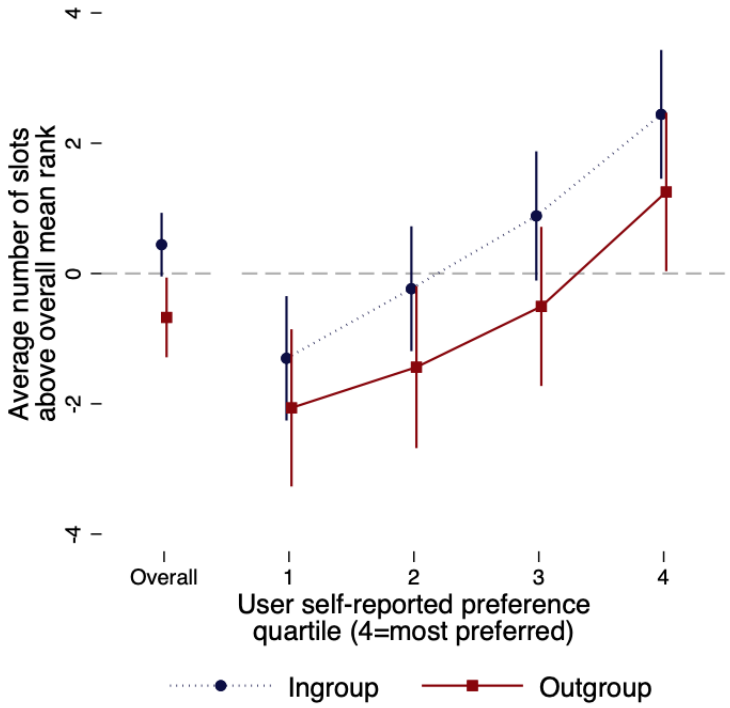

Figure 3.6. Relationship between NF Algorithmic Ranking and In-group Status Conditional on Subject Explicit Preference - India. We show the mean ranking of in-group and out-group posts above the overall mean, and then we show this by subject stated preference. The normalized subject explicit preference quartile is the across subject quartile of within subject z-scores for stated preference. Each subjects ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins. Ingroup is defined as same religion.

In contrast (and again consistent with the findings in the US data), we see little difference in the algorithm's rankings for ingroup versus outgroup recommendations with PYMK. Given our smaller sample size for India relative to the US, our PYMK results are somewhat noisier (our 95% confidence interval does not let us rule out an own-group versus out-group difference in rankings less than 0.839 slots), but the point estiamte is quite small, about one-fifth of a ranking slot.

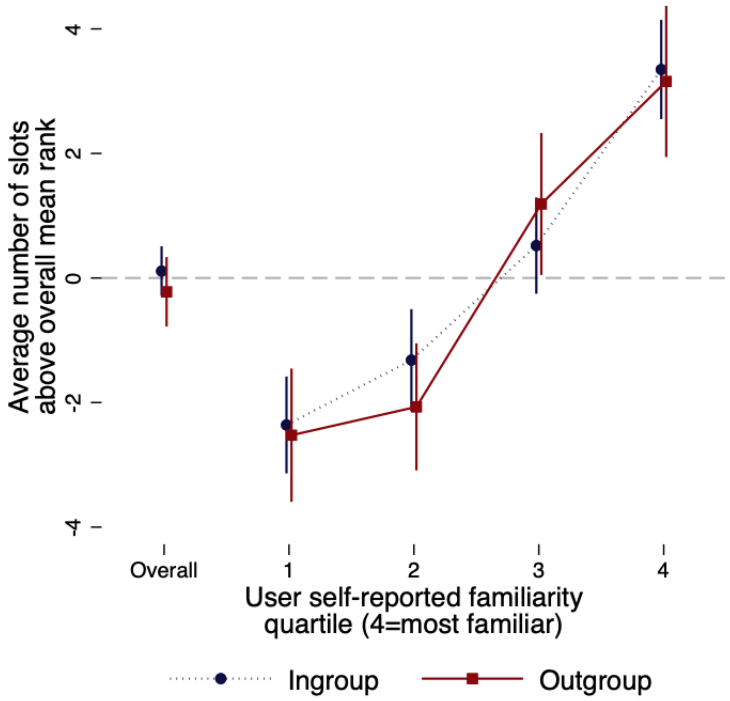

Figure 3.7. Relationship between PYMK Algorithmic Ranking and In-group Status Conditional on Subject Explicit Preference - India. We show the mean ranking of in-group and out-group recommendations above the overall mean, and then we show this by subject stated familiarity. The normalized subject explicit familiarity quartile is the across subject quartile of within subject z-scores for stated preference for a familiarity with a suggested friend. Each subjects ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins. Ingroup is defined as same religion.

And as with the US data, even though we see a difference in Newsfeed rankings of own-group versus out-group content in India, we do not see much detectable bias in explicit user preferences.

The tendency of algorithms to learn our preferences from our worst selves does not appear to be specific to any particular type of bias (race versus religion versus other salient in-group vs. out-group distinctions), or to any particular setting or country context.

3.5 Lab Experiments

The comparison of bias between the Newsfeed and PYMK algorithms is obviously not a perfectly controlled experiment. While the decision-making contexts differ between Newsfeed and PYMK in terms of how automatic versus deliberate users are when interacting with the algorithms, other things also differ across the algorithms.

To address the problem of confounding we carry out a series of carefully controlled laboratory experiments, which mimic the key features of online settings where choices are algorithmically curated for users. These experiments have the advantage of more precisely isolating and testing the implications of our model, although of course at the cost of substituting behavior in the lab setting rather than in the real world. We first carry out two lab experiments that show how more automatic decision-making contexts exacerbate the problem of bias, which in turn introduces more bias to algorithms built using data from quicker, more automatic decisions. We then carry out a third experiment that shows how the rankings from such an algorithm create a "double penalty" for out-group content relative to situations where subjects choose from randomly-ranked content.

3.5.1 Experiments 1 and 2 Design

Our lab experiments mimic the task of an algorithm that must rank content for a user based on choices from past decisions (potentially of the same user and also other users). The advantage of these lab experiments is the ability to randomize subjects to both own-group vs. out-group content, and to rushed versus non-rushed decision-making contexts. And, because we build the choice-curating algorithm ourselves using the responses from our own lab study subjects as the training data, we know exactly how the algorithms were constructed (unlike with Facebook).

Our subjects look at a series of movie recommendations on a screen, where each movie recommendation is attached to a poster by name ("Amanda A. recommends..."), and choose

amongst a subset of those movies to potentially watch. The movies and posters are shown three at a time on the screen, much like results from search algorithms or social media sites might show up. The respondent can click on a button to read a fuller review from the recommender. This is similar to many actual online environments where choices are curated, users see limited information about the content, see a person attached to that content (Twitter handle/twitter post, Facebook post/name, etc.), and are able to click to get a little more information before making a decision about whether to engage with the content (like select a movie). We selected 42 movies total from various genres. Reviews are taken from the public dataset used in Maas et al. (2011).[12] The respondent is instructed to choose 4 movies, and told that once chosen they would get a link to watch one of their chosen movies. Since the recommendations show up 3 at a time, the respondent needs to click on "load more" at least once to choose 4 movies.

We use an audit-study design that randomly assigns names to movie reviews, selecting names of the recommenders or posters to saliently signal race and gender. We draw the list of names from previous studies that have used a similar design to understand different economically-meaningful decisions like hiring (Bertrand and Mullainathan 2004, Agan and Starr 2018, Milkman, Akinola, and Chugh 2012). The distribution of race among our fictional "posters" is designed to mimic the U.S. population overall.[13] Randomization of names means that we are holding constant true underlying preference for the movies themselves.

We also randomize subjects to make decisions under one of two different choice contexts:

- A *rushed* condition in which the subject is told they will have 5 minutes to make their selections, and are told that this is not much time given the task at hand. To reinforce this, the clearly visible countdown clock on the left of the screen counts in milliseconds

12. Additional information about the movies, such as pictures to attach to the movies and ratings, were taking from IMDB.com

13. In expectation a user would see 64% white-signalling names, 22% Hispanic signalling names, and 14% Black-signalling names.

so that the countdown is moving quickly while the respondent is deciding.

- A *non-rushed* condition in which the respondent is told they have 15 minutes to decide, that this is plenty of time, and the timer counts in minutes, moving much more slowly than in the "rushed" condition.

We carried out two versions of the lab experiment:

- *Experiment 1* involved $N = 981$ study subjects recruited through the Prolific platform, with a sample intended to be representative of the US population with respect to race.
- *Experiment 2* involved $N = 753$ study subjects from Prolific who were white males, a design modification intended to make the job of defining in-group versus out-group posts (one of the key features of our experiment) easier.

Experiment 1's sample is roughly representative of the country as a whole with respect to race and ethnicity (63% white, 23% Hispanic, 14% Black), slightly over-represents females (61%), is more highly educated than the population overall (32% have a Bachelor's degree and 31% have more than a Bachelor's degree), and has an average age of 29.6. The study sample for Experiment 2 is older (38.8 years of age) with lower levels of schooling. We have a total of $N = 4,859$ movie selections from Experiment 1 and $N = 3,001$ from Experiment 2.

3.5.2 Experiments 1 and 2 Results

The results for study subject choices are shown in Figures 3.8 and 3.9. In the deliberate decision-making condition in Experiment 1, subjects are 1.4 percentage points more likely to choose movies recommended by an own-group member, which represents a 7.8% increase over the mean click-rate for outgroup recommended movies of 17.8 ($p < 0.10$). But when rushed, subjects are 3.1 percentage points more likely to choose a movie recommended by

an own-group poster ($p < 0.01$; an 18% increase over the mean click-rate for out-group recommended movies). In Experiment 2 the own-group preference is equal to 0.8 pp in the deliberate condition (6% higher than the out-group mean click rate) and 0.5 pp (12%) in the rushed condition. Neither is statistically significant, but our standard errors here do not allow us to rule out own-group favoritism as large as 2.5 pp in the deliberate condition and 2.2 pp in the rushed condition.

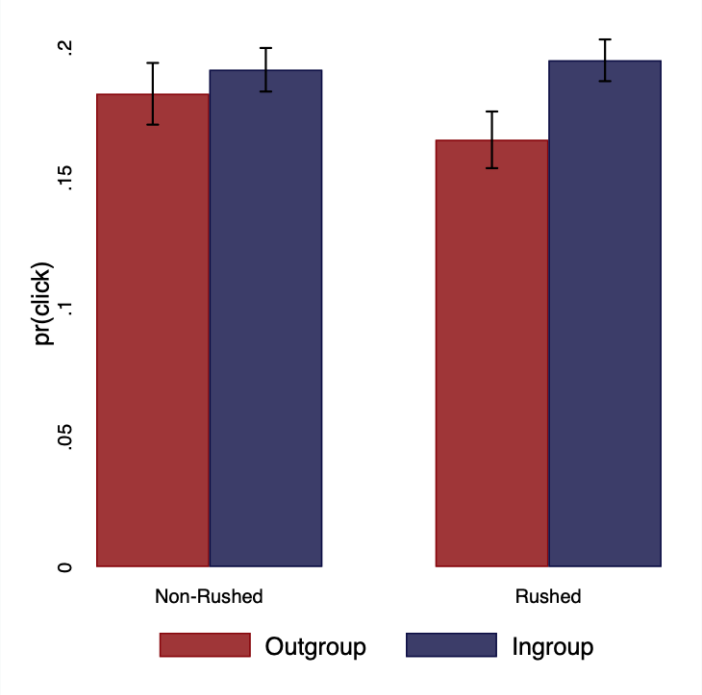

Figure 3.8. Probability Choose Movie by Treatment and Recommender Type: Experiment 1. This graph shows the engagement patterns for in-group vs. out-group recommended-content, by rushed and non-rushed conditions.

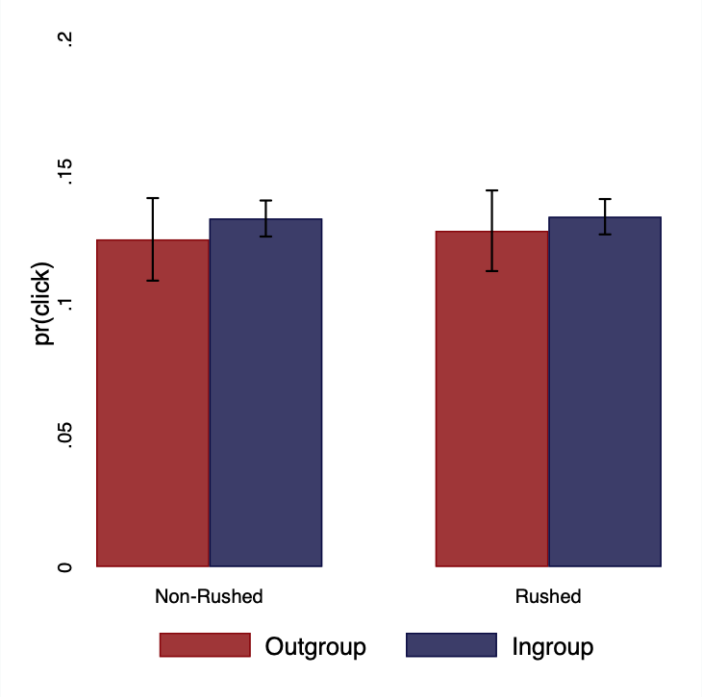

Figure 3.9. Probability Choose Movie by Treatment and Recommender Type: Experiment 2. This graph shows the engagement patterns for in-group vs. out-group recommended-content, by rushed and non-rushed conditions.

We then fed the results of our two lab experiments into an algorithm-construction pipeline. Specifically, for each study sample separately, we build a machine learning algorithm that takes the subject choices from a given experimental condition (rushed versus non-rushed decision context) within a given lab experiment, then predict user selections. We estimate a random forest, a widely-used classification algorithm, with inputs given by features of both the movie options and the study subjects. The output of this algorithm is a rank-ordering of movie choices by the predicted likelihood of user selection. This is basically how actual curation algorithms work, with the only key differences being our dataset is much smaller than what is used with commercial algorithms.

The top right panel of Figure 3.10 shows how the algorithm sorts movie recommendations for subjects in Experiment 1, when trained separately on choices from the rushed and non-rushed conditions. The vertical axis depicts where each movie gets ranked in terms of number of slots relative to the average movie ranking. Own-group posts end up more towards the top

when the algorithm is trained on data from the rushed condition, equal to 2.7 ranking slots above those from outgroup posts ($p < .05$), while from the non-rushed condition in-group recommendations only move about 1 ranking slots above outgroup posts ($p < 0.05$).

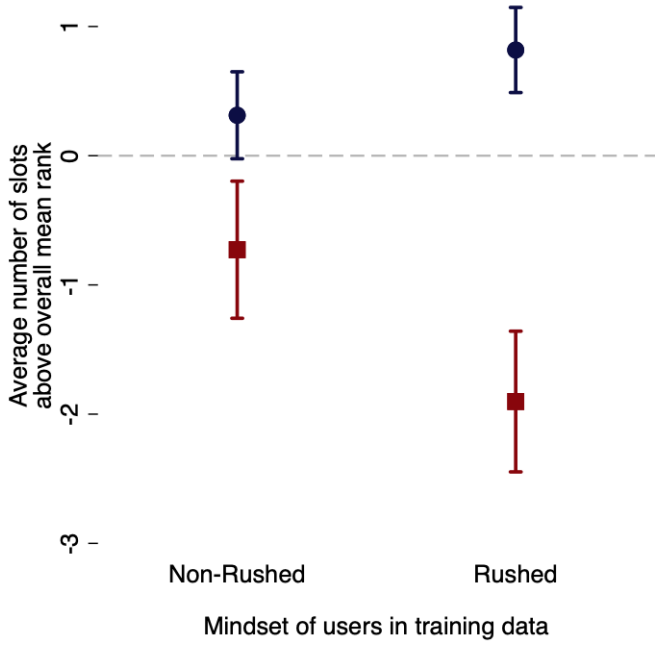

Figure 3.10. Algorithmic Ranking Trained on Experiment 1 Choices. This graph shows the results of using the data from the rushed and non-rushed conditions separately to build two separate algorithmic predictions that rank content by subject engagement choices.

For Experiment 2, even though the in-group favoritism is not statistically significant on average for the rushed or deliberate conditions, as the model predicts when we feed these choices into our algorithm to sort movie recommendations we see very similar results to Experiment 1. There is up-ranking of own-group posts in the rushed condition (3.3 ranking slots above outgroup, $p < .05$) and 1.9 ranking slots above outgroup in the deliberate condition ($p < 0.05$). We return to the issue of how and why algorithms can find, and recreate, own-group bias even when it is not detectable in the raw data from people's choices in the discussion section below.

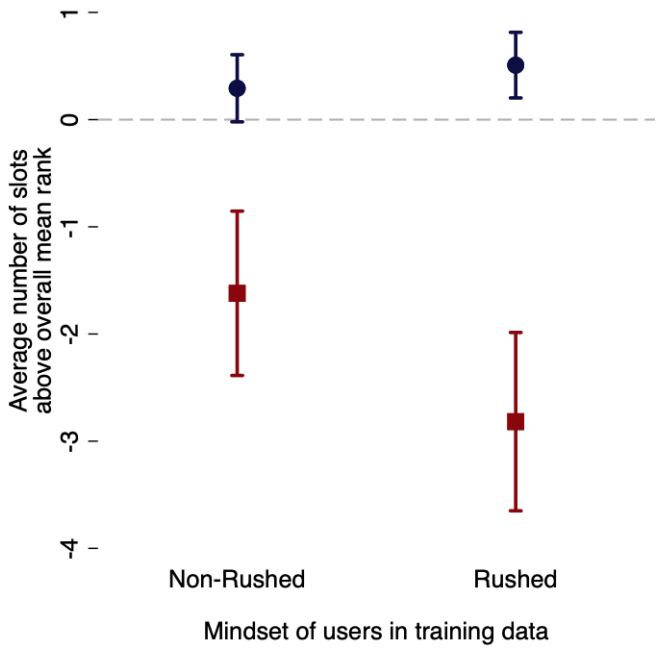

Figure 3.11. Algorithmic Ranking Trained on Experiment 2 Choices. This graph shows the results of using the data from the rushed and non-rushed conditions separately to build two separate algorithmic predictions that rank content by subject engagement choices.

3.5.3 Experiment 3

Experiments 1 and 2 demonstrate that algorithms up-rank own-group content, particularly when choices come from contexts where decision-making is done in a rushed context and so more likely to be automatic rather than deliberate. Our final lab experiment provides at least suggestive evidence that the use of such an algorithm to curate choice options for subjects creates a "double penalty" against out-group content, particularly in settings of rushed decision-making.

For Experiment 3 we enrolled a total of (N=757) US white male study subjects on the Prolific platform. We replicated the rushed condition from Experiments 1 and 2, but now randomized subjects to two conditions:

- A *randomly-ranked* condition in which subjects are shown candidate moving recommendations that are randomly ranked.

- An *algorithmically-ranked* condition in which subjects are shown movie recommendations that are ranked on the screen using the algorithm that we built using data from the rushed condition for Experiment 2, which, as shown above, up-ranks in-group recommendations and down-ranks out-group recommendations.

Figure 3.12 shows that random ranking, as expected, has in-group and out-group recommended content ranked very similarly on average (the difference is −0.10 slots, standard error 0.27). In contrast, we can see on the right figure that the curation algorithm built using data from the "rushed" subjects in Lab experiment 2 substantially up-ranks posts from own-group recommenders versus out-group recommenders (the difference is 8.30 ranking slots, standard error 0.31). By way of comparison, this is about as large an impact on the algorithm's rankings as the movie genre being the study subject's favorite genre (equal to 8.43 ranking slots, standard error 0.22).

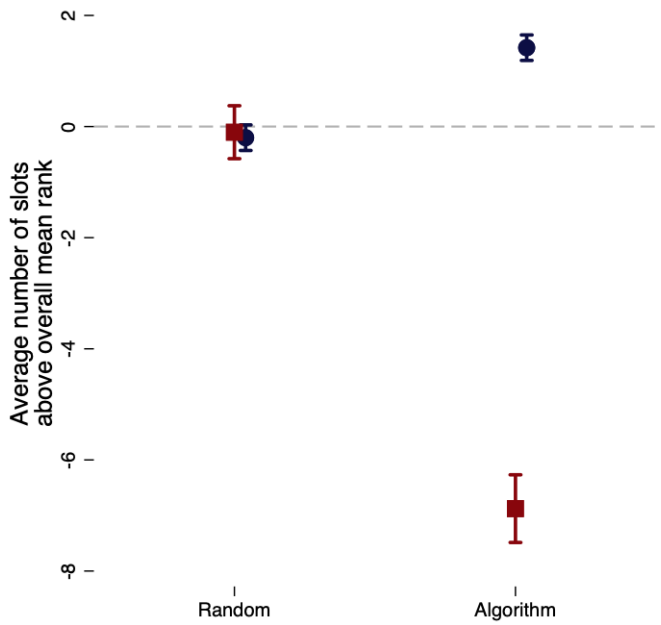

Figure 3.12. Ranking of posts in Random vs. Algorithmic Ranking Treatments.

Figure 3.13 shows that when choices are randomly ranked there is a 0.9 pp difference

in favor of in-group content that is not statistically significant (standard error of 0.8 pp), which stems from the single penalty of implicit bias. In contrast, when we add on the double penalty of the curation algorithm down-ranking out-group content, the in-group vs. out-group difference in the chances subjects engage with the content increases to 3.6 pp (standard error 1.3 pp). While these data are from the lab setting, not real-world choices, the results are at least consistent with the idea that algorithms, especially in rushed decision-making contexts, can exacerbate the problem of people's own biases by reducing the chances people see out-group content—thereby creating a "double penalty" against out-group members.

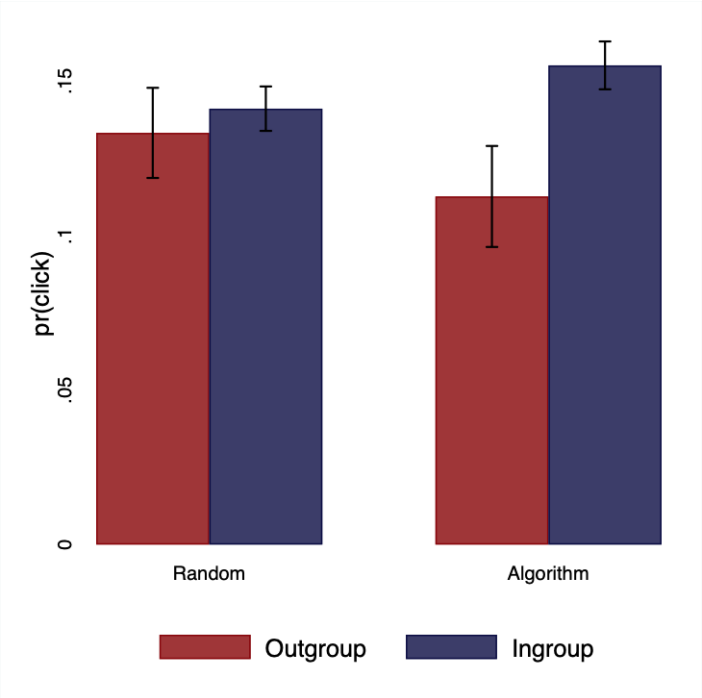

Figure 3.13. Probability Choose Movie by Treatment and Recommender Type: Experiment 3.

3.6 Discussion

Modern life increasingly involves algorithms designed to predict our preferences and help us make choices. Algorithm designers apply industrial-strength machine learning techniques to the data to wring as much signal as possible about the behavioral outcomes captured by

the data. But remarkably little attention is devoted to the possibility that the behavioral measures captured in these data might diverge from our preferences. The attention devoted to the machine learning engineering of these algorithms is not matched by similar attention to understanding the psychology of the human beings whose behavior generates the data.

One of the major advances in psychology over the past several decades is the development of "dual-system" theories of cognition, which help us see why our behaviors may not always reflect our true underlying preferences.[14] Because deliberate conscious thoughts (what psychologists sometimes call "system 2 thinking,") is mentally taxing, we all rely as much as possible on fast, easy automatic responses ("system 1") in ways that we are typically not even aware of.

These implicit cognitions can sometimes be in direct conflict with our conscious thoughts. System 1 cognition is designed to be fast and easy, not necessarily accurate. We sometimes make snap judgments about correlations between people's behavior and demographic characteristics like sex, age and race—that is, we might stereotype, even if only subconsciously.[15] We often fail to process all the information that is available in ways that might adversely affect out-group members, or we might interpret information inconsistently leading to inconsistencies (or "noise") in our behaviors (Kahneman, Sibony, and Sunstein, 2021).

There are two key implications of the findings reported here, one practical and the other conceptual.

The practical implication here stems from evidence of bias in one of the most widely-used social media algorithms in the world, which sorts Facebook's Newsfeed posts. Given

14. The literature on dual-system theories is vast. For an excellent and influential summary of this literature in psychology, see Kahneman (2011). Within economics, models of dual systems thinking include Cunningham (2013) and Fudenberg and Levine (2006).

15. A great deal of attention has been devoted to ways of measuring implicit biases, such as the implicit association test or IAT (Greenwald, McGhee, and Schwartz, 1998; Bertrand, Chugh, and Mullainathan, 2005). One challenge is that we only have IAT tests for convenient, rather than representative, samples of volunteers (Arkes and Tetlock, 2004). But implicit bias seems to be prevalent among those who volunteer to take the tests, and some people argue that it is correlated with actual behavior (Hehman, Flake, and Calanchini, 2018; Orchard and Price, 2017).

the growing role of curation algorithms in shaping people's choice sets more generally, the potential for social harm both now and in the future seems enormous. These disparities are especially troubling because algorithms are not simply curating products and services; they are matchmakers mediating how people connect with each other. As a result of algorithmic ranking based on choices, users have their own biases magnified in what they see, but marginalized users also have to work twice as hard to have their content seen by others.

The second implication is conceptual. We have argued, and demonstrated in a lab setting, that algorithms have the potential not just to mirror our explicit biases, but exaggerate our biases beyond what our deliberate selves would like by mistaking our implicit biases for our actual preferences. Algorithms are also problematic in another way. When people recognize a gap between their preferences and their choices, they can do something about it. They can adjust their environments and how they approach the choice to bring these two in line with each other. But the algorithm blindly learns from behavior without access to preferences. Since it, not us, constructs the choice set, it impedes the opportunity for us to learn the problem, which can make the gap more persistent.

CHAPTER 4

MOTIVATED AND SELECTIVE ATTENTION: EVIDENCE FROM NEW JERSEY POLICE OFFICERS

4.1 Introduction

When do we seek information to help us make a decision and what exactly do we do with that information? Many fields in economics have deviated from assumptions of perfectly informed and fully attentive agents, but relatively little empirical work has demonstrated how, when, and why people acquire *useful* information. This paper fills that gap by providing evidence on endogenous information acquisition and belief updating in an especially high-stakes policy setting where information is literally a push of a button away.

In New Jersey, when one is arrested, bail does not determine whether you wait in jail or remain free during the course of your criminal proceedings. Instead, the law mandates that (only) two specific risks be taken into account to determine whether one should be detained: (1) risk of failing to appear for court dates and (2) risk of committing crimes while released. As a result, judges are given access to an algorithmic risk tool to help make predictions and assessments of the two risks. Importantly, before a defendant ever appears before a judge, a police officer involved in the defendant's arrest and booking makes a similar decision, deciding whether a *warrant* should be issued—which results in the defendant remaining in custody until seeing a judge—or a *summons* should be issued, which results in the defendant remaining free until seeing a judge *and* precludes the judge from detaining the defendant at the intial hearing. Therefore, judges are the final determinants, but the initial choices by the officers have a significant impact on the choice set of the judge. To aid collaboration between judges, attorneys, and officers, the officers also have access to the algortihmic recommendations that will be available to the judge. Officers have complete discretion over whether to click the button to view the algortihmic recommendation and,

for a select category of charges, they have complete discretion over whether to follow the recommendation if they do see it. I focus on these two forms of discretion and study how officers choose to consult the algorithmic risk scores and how those choices affect the warrant decision. See Section 4.2 for further background and institutional details.

To answer these questions, I leverage a unique dataset that captures the universe of arrests in NJ from 2018 to 2021. For each arrest, I observe the details of the arrest (e.g., charge, circumstance, demographics), whether the officer consulted the algorithm, the algorithm's prediction, and the ultimate warrant recommendation of the officer. Importantly, in the data I observe the algorithm's prediction regardless of whether the officer consulted the algorithm. To my knowledge, this is the first detailed view (exact data on what officers see/do and behind-the-scenes info on the algorithm) into how people make use of the algorithms available to them. I provide further information on the data and my empirical framework in Section 4.3.

There is overwhelming evidence that officers don't behave strictly according to the law set out by the state, which requires that they only consider the two future risks. Instead they seem to have a rich set of private preferences that influence both their information acquisition as well as their eventual warrant decisions. First, in Section 4.4, I provide descriptive evidence suggesting that officers' behavior is highly influenced by the current charge, despite the fact that current charge is not associated with current risk. This pattern represents a deviation from the state's objective function, which prioritizes the two future risks, but it is difficult to normatively judge this influence as it requires taking a stance on potentially complex considerations such as fairness and different valuations of different crimes.

In section 4.5.1, I document the influence of another factor that is straightforward to judge on normative and legal grounds: race. I provide evidence for several forms of race based discrimination. First officers are *more* likely to consult risk predictions for minority defendants. Second, they are *more* likely to issue a warrant for minority defendants,

controlling for a battery mitigating factors. Third, the first two findings are linked by the fact that officers are less influenced by the risk scores for minority defendants. In short, when it comes to racial minorities, officers are more likely to look for signals of risk, less likely to believe the signals they see and are thus more likely to recommend detention.

This paper contributes to several literatures. First, a vast collection of papers across economics and psychology has explored how people pay **attention to information**. The research can study attention in two senses, as clarified by Gabaix (2019)—how do people acquire information and how do they incorporate information that they have acquired (or have been given)? For example, much has been learned about people's preferences (e.g., Card et al. 2012, Perez-Truglia and Troiano 2018, Allcott 2011), beliefs (e.g., Bursztyn, González, and Yanagizawa-Drott 2020), and technological/decision-making constraints (e.g., Karlan et al. 2016) by exogenously providing information. In another stream, others have studied how people's decision to acquire information (that may or may not have instrumental value) is affected by those channels (For examples, see Golman, Hagmann, and Loewenstein 2017 for a useful review). I connect these two streams by showing a pattern of acquisition and updating that is most consistent with a motivated account. Officers are more likely to consult information for certain groups but also less likely to believe the information when they see it, which suggests the lengths people may go to in order to manage congruence with their private beliefs, outside information, and observable behaviors.

Second, the persistent racial disparities in attention to information contributes to our understanding of the **economics of discrimination**. The key tradition in this literature has been documenting and distinguishing between taste-based and (accurate or inaccurate) statistical discrimination (Becker 1957, Phelps 1972, Arrow 1973, Aigner and Cain 1977, Bohren et al. 2019). (See Bertrand and Duflo (2017) for a useful review of many of the central empirical contributions to this literature.) Most closely related to the present research, Bartoš et al. (2016) study *attention discrimination* in rental and labor markets and provide

evidence from field experiments that screeners seek out *more* information for minority groups as the market becomes less selective. My findings are consistent with such an endogenous attention account. In a very different market (jail recommendation decisions) in observational data I demonstrate that officers appear to strategically use information to facilitate their discrimination. These findings call for a broader rethinking of interventions when agents with racial animus also face information asymmetries.

Next, studying how officers navigate a prediction policy problem (Kleinberg et al. 2015) with an algorithmic risk tool contributes to the literature on **human plus machine decision making**. Previous work in this area has documented that algorithms of varying sophistication can outperform humans in a wide variety of tasks (Yeomans et al. 2019, Kleinberg et al. 2018b, Meehl 1954, Dawes, Faust, and Meehl 1989). This theme has given rise to a subliterature exploring an innate human tendency to seek out or avoid relying on algorithms (Dietvorst, Simmons, and Massey 2015, Dietvorst, Simmons, and Massey 2018, Logg, Minson, and Moore 2019). Meanwhile, other papers have documented the bias that algorithms themselves may introduce to a setting. (Obermeyer et al. 2019, Buolamwini and Gebru 2018, Corbett-Davies et al. 2017) This paper ties these strands together and is the first to document how humans interact with an algorithm in a high-stakes field setting. I show that while human+algorithm (human in the loop) decision-making offers large potential for synergy between man and machine in principle, it also introduces enormous scope for manipulation and discriminatory overriding.

4.2 Institutional Background

Cash bail is a fraught feature of the American criminal justice system. Historically the system has kept people in jail due to their inability to pay, instead of the threat they pose to public safety (new criminal activity) or integrity of courts (failure to appear). Several states are turning to reforms to reduce unnecessary detention.

4.2.1 New Jersey Reform

As of 2017, New Jersey has abolished cash bail in favor of a series of sweeping policy changes frequently referred to as "Criminal Justice Reform" (CJR). There are two key aspects of the reform that are central to the present research — presumption of cases and risk tool assessments — which I discuss in the following sections.

4.2.2 Discretion and eligible cases

After an arrest, once probable cause has been established, a judicial officer determines which of the following categories the most severe of the alleged crimes falls into: mandatory warrant, presumed warrant, or presumed summons. The mandatory warrant cases include murder, sexual assault, and carjacking. For these cases there is no officer discretion and a warrant must be issued. For all other cases, there is a default statutory recommendation but ultimately the officer can decide freely whether to issue a warrant without explanation. For ease of interpretation I focus exclusively on cases that are presumed summons. To overcome the summons presumption, officers are statutorily recommended to consider several factors, including the "Public Safety Assessment" risk score, the presence of unclassifiable charges, out of state criminal history, and any extenuating juvenile or domestic violence considerations. Note that none of these are strictly enforced.

4.2.3 Public safety risk and assessment

The New Jersey "Public Safety Assessment" (PSA) is an algorithmic risk prediction tool and decision aid built using over 1.5 million cases from over 300 jurisdictions across the US. It is designed to predict two independent risk scores each of which ranges from 1 to 6 (integer-valued) where 6 is the highest risk[1]: risk of failing to appear for court dates and

1. There is also a third risk score for new *violent* criminal activity, but I only consider cases where this score is not a relevant factor.

risk of engaging in new criminal activity during criminal proceedings for the current case.

When calculating risk scores, the algorithm considers nine factors:

1. Age at current arrest
2. Violent offense
3. Pending charges at time of arrest
4. Prior misdemeanor
5. Prior felony
6. Prior failure to appear in past two years
7. Prior failure to appear over two years ago
8. Previously incarcerated

4.3 Framework and Data

4.3.1 A simple model

To motivate and structure the empirical strategy, I consider a stylized model where an officer arrests a defendant and makes two binary decisions: run the algorithm (r) and issue a warrant (w).

Suppose the risk that the defendant with characteristics (x, z)—where x is seen by the officer and the researcher, while z is seen only by the officer—poses to society is captured by $t \in \{H, L\}$. The state expects the officer to issue a warrant based on her beliefs about t. The state gives officer optional access to an algorithm to help form beliefs about t. The following game ensues:

Timing

1. (Nature chooses $(x, z, t)_i$ from some joint distribution)

2. Officer j sees $(x, z)_i$ and forms beliefs; An algorithm, π(x), can generate a signal of risk $s \in \{h, l\}$.

3. Officer *chooses* to observe s; If s is observed $r = 1$, otherwise $r = 0$.

   (a) Officer subjectively updates beliefs, possibly according to Bayes rule or some heuristic

4. Officer chooses whether to issue warrant, w (and payoffs are realized) [2]

What can we learn about the preferences of the officer based on the profile of actions (r, w)? First, consider two defendants, M and F. Let E$[r_M]$ capture the probability that the officer observes the algorithm for defendant M and E$[r_F]$ is the same quantity for defendant F. Then if E$[r_M]$ > E$[r_F]$, this implies that the value of the signal that the algortihm produces is higher for the defendant M than it is for defendant F, a result that has been shown as early as Blackwell (1953).

Two forms of value are captured by the differential propensity to run the algorithm: instrumental and non-instrumental. Officers may be more likely to observe the algorithm for defendant M because they believe the result will be more informative or they may have some other goal, such as signaling to their superiors or to themselves.

Therefore, if we see differential rates of observing the algortihm, we can turn to the influence of the algortihm on behavior to identify what kind of value (instrumental or non-instrumental) the algortihm presents to the officer. If the officer is observing the algortihm because she beleives it is more informative, then it should be weakly more influential on

2. Note distinction between the officer's payoff function and the social payoff function. Note that the social planner stands to gain if there is information in Z that is valuable for predicting type, but stands to lose if there is information in Z that is salient and persuasive to the officer, but is not predictive of type. Whether or not there is information in Z, the algorithm would prefer for the defendant to not use X, since it has built at minimum the BLP of R given X, but perhaps has a more flexible function such that the officer cannot hope to outperform.

beliefs and therefore weakly more influential on warrant behavior. If instead we see that the algorithm more likely to be observed for defendant M, but the influence of the algorithm is lower for defendant M, that is evidence of running the algorithm to achieve some motivated goal (Kunda 1990).

4.3.2 Data

To empirically assess the underlying motivations of officers I analyze data from the universe of over 217,000 arrests in the state of New Jersey from 2018 to 2020. For each arrest I observe the details of the charge, the race and gender of the defendant, the algorithmic risk score for the defendant, whether the officer saw the risk score, and whether the officer issued a warrant. Importantly, I observe the risk score for all defendants, regardless of whether the officer chose to see it. Table 4.1 shows basic summary statistics of the data.

4.4 Descriptive Evidence

I begin by documenting basic descriptive facts about the algorithm itself and officers' use of it.

As indicated in Section 4.2 the PSA generates two scores. Figure 4.1 shows the intuitive result that the two scores are highly correlated.

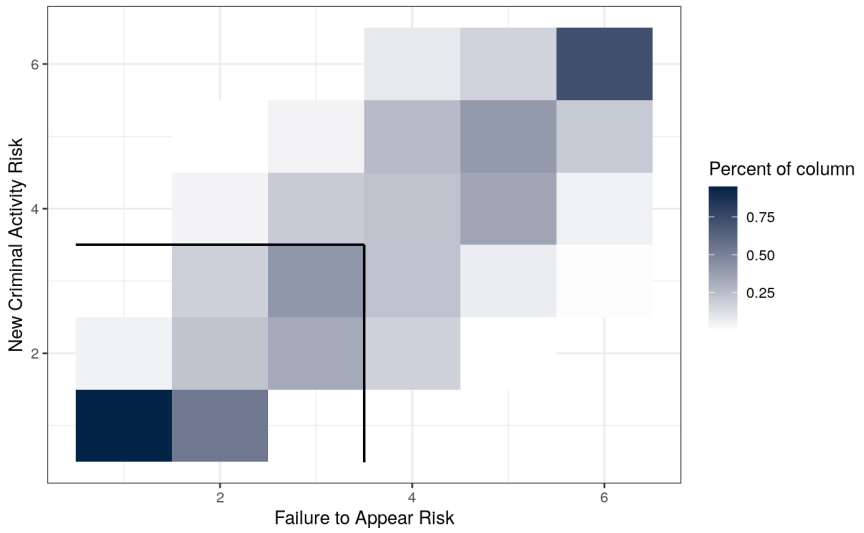

Figure 4.1. Relationship between risk scores This is a graph of the joint distribution of two algorithmic risk prediction scores: failure to appear for court appearances (FTA) and new criminal activity if released (NCA). Each square represents the percentage of defendants with a given FTA score that have a particular NCA score. All columns sum to 1. The black line represents the threshold for recommending a warrant—if either score is above a 3, a warrant is recommended by the algorithm.

As documented in other jurisdictions (e.g., Angwin et al. 2016) and settings (e.g., Buolamwini and Gebru 2018), Figure 4.2 shows that there are noticeable differences in the algorithmic risk profiles across races.

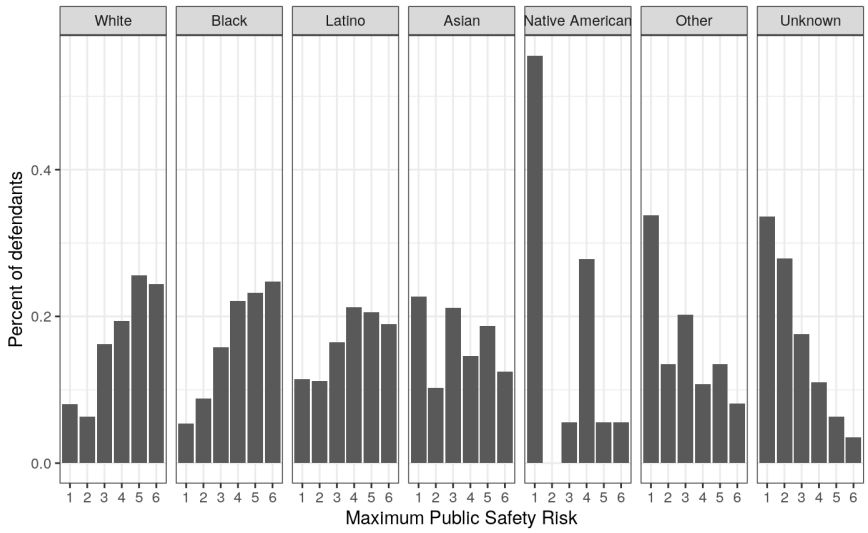

Figure 4.2. Risk score distribution by race This graph shows the distribution of the max of two algorithmic risk scores (failure to appear and new criminal activity) by race.

A simple decision rule that each officer may implement for the PSA is to determine whether it is valuable or not, in general. In this case, some officers would always run the PSA, while others would never run it. Instead, what we see in Figure 4.3, is that officers vary enormously *from case to case*. Understanding the causes and consequences of this variation is a central goal for this paper.

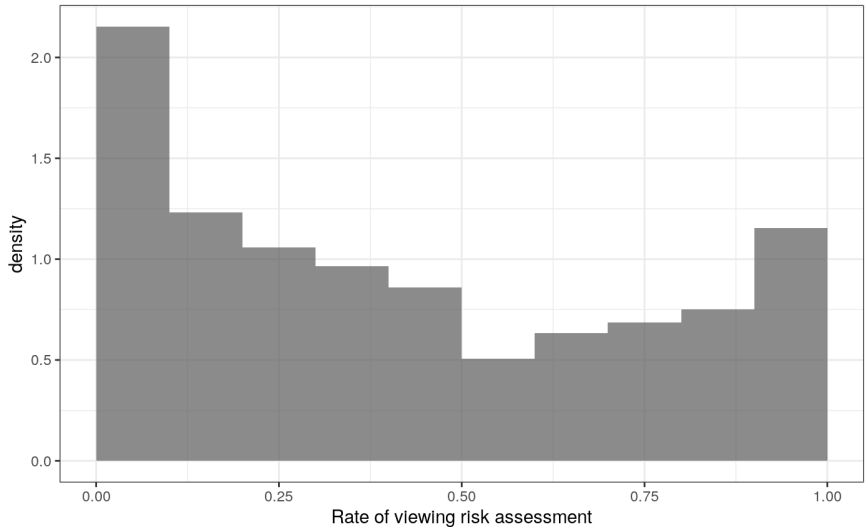

Figure 4.3. Variation in information acquisition This graph shows the distribution of the rate at which officers choose to view the algorithmic risk scores. Only officers who are associated with at least 10 arrests are included.

Consistent with the "honest strategy" from the framework in section 4.3, officers could seek out information when they have less certainty regarding the defendant's risk. Figure 4.4, suggests that this is not the case since there is no clear relationship between the risk of the defendant and the probability of seeking out information.

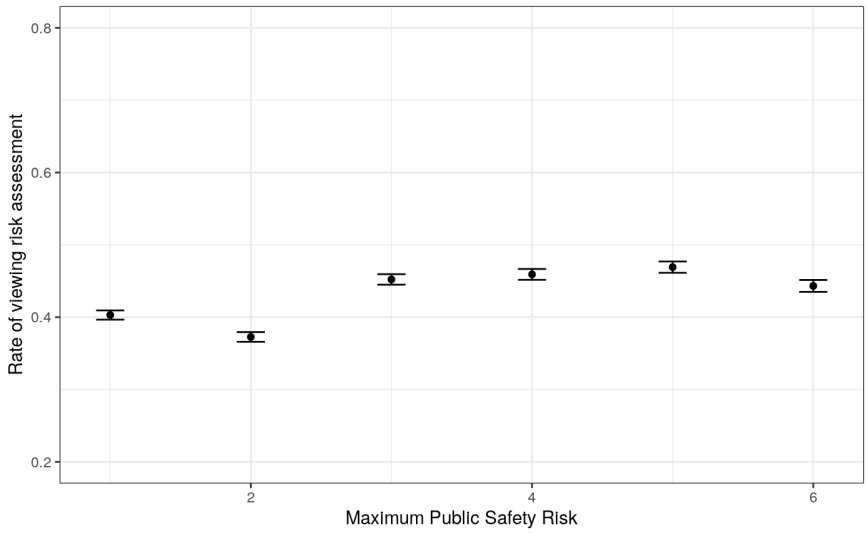

Figure 4.4. Information acquisition by algorithmic risk This graph shows the rate at which officers choose to view the algorithmic risk score for defendants at each level of overall public safety. Overall public safety is defined by the max of two algorithmic risk scores (failure to appear and new criminal activity).

Figure 4.5 disaggregates figure 4.4 by the severity of the current charge. It shows that charge severity is a first-order determinant of running the PSA while risk (though not necessarily uncertainty) also has some influence albeit much less.

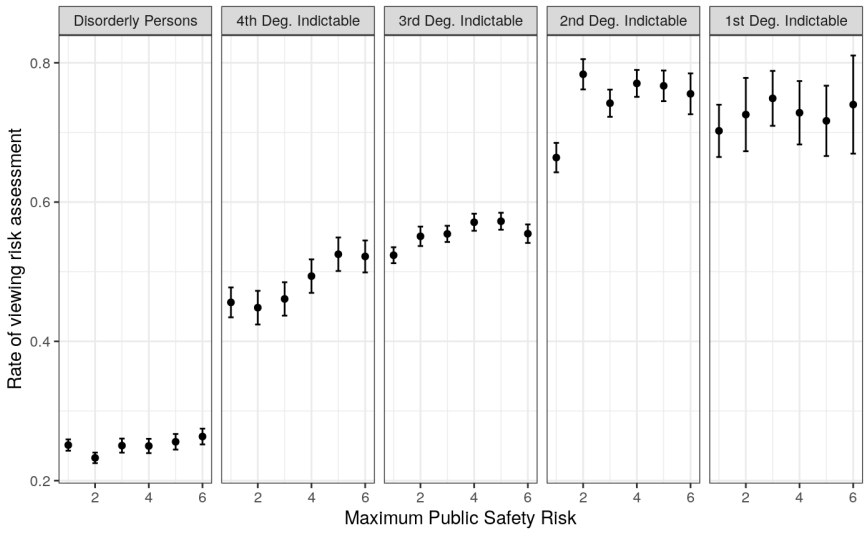

Figure 4.5. Information acquisition by algorithmic risk and current charge This graph shows the rate at which officers choose to view the algorithmic risk score for defendants at each level of overall public safety. Overall public safety risk is defined by the max of two algorithmic risk scores (failure to appear and new criminal activity). Each panel is a different classification of crime—classifications are ordered from left to right by increasing severity.

What then is the influence of running the PSA on warrant behavior? Figure 4.6 provides at least suggestive evidence that the PSA matters a lot. When officers don't see it, there is practically no relationship between risk and warrants, but when it is run the relationship strikingly large. [3]

3. If relying strictly on Figure 4.6, this conclusion should be taken with some caution because officer selection on prior beliefs and predisposition could generate this data without any causal influence of the algorithm. However, I provide empirical evidence that the algorithm does have a causal effect on warrant behavior through several strategies in omitted supplemental work. Those estimates suggest that the influence of selection on the relative slopes of the two panels in Figure 4.6 is minimal.

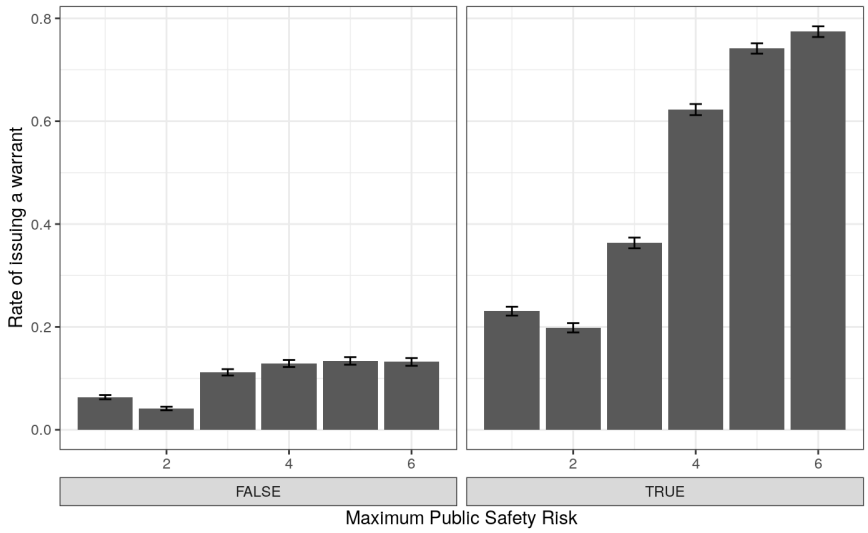

Figure 4.6. Warrant behavior and information acquisition This graph shows the rate at which an officer issues a warrant for defendants at each level of overall public safety. Overall public safety risk is defined by the max of two algorithmic risk scores (failure to appear and new criminal activity). The left panel shows the relationship when officers do not view the algorithm. The right panel shows the relationship when they do.

That the officers select on charge severity could be efficient if severity were correlated with relevant moments of the risk distribution. I reject this hypothesis in Figures 4.7 and [insert graph of the variance] by showing that the mean risk does not vary systematically with charge nor does the variance.

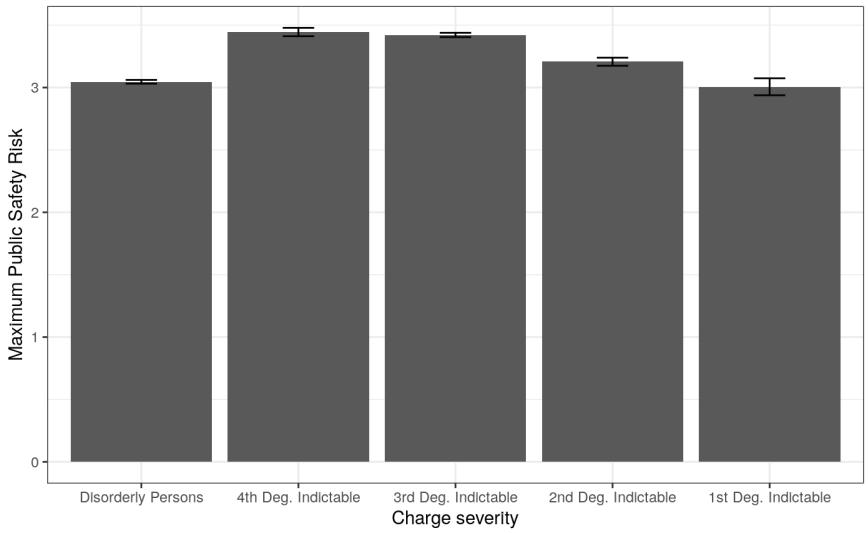

Figure 4.7. Current charge and algorithmic risk This graph shows the mean overall future public safety risk by the severity of the current accused offense. Overall future public safety risk is defined by the max of two algorithmic risk scores (failure to appear and new criminal activity). Current charge classifications are ordered from left to right by increasing severity.

Despite the independence between risk and charge severity, officers seem to heavily prioritize charge severity when issuing warrants. Figure 4.8 shows that those charged with the most severe charges (1st degree indictables) are nearly six times more likely to be issued warrants than those with the least severe charges (Disorderly Persons), despite the fact that they do not pose any more risk on average.

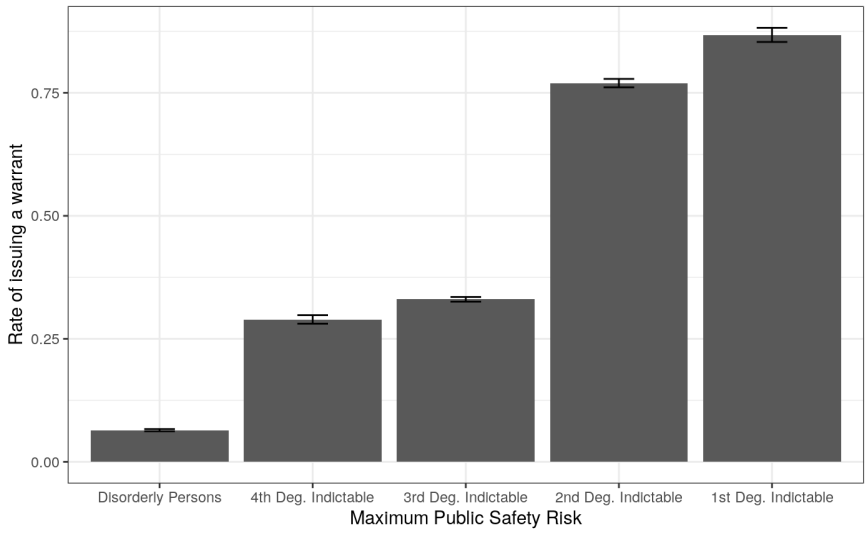

Figure 4.8. Warrant and current charge This graph shows the rate at which warrants are issued by the severity of the current accused offense. Overall future public safety risk is defined by the max of two algorithmic risk scores (failure to appear and new criminal activity). Current charge classifications are ordered from left to right by increasing severity.

Figure 4.9 makes this point more directly. For DP defendants who are high risk (ie, the algorithm recommends a warrant), they are only issued warrants x% of the time when's the algorithm is not run and just over 50% when the algorithm is run. On the other hand, for high risk first degree indictable defendants those rates are 55% and 90% when algorithm is not and is run respectively.

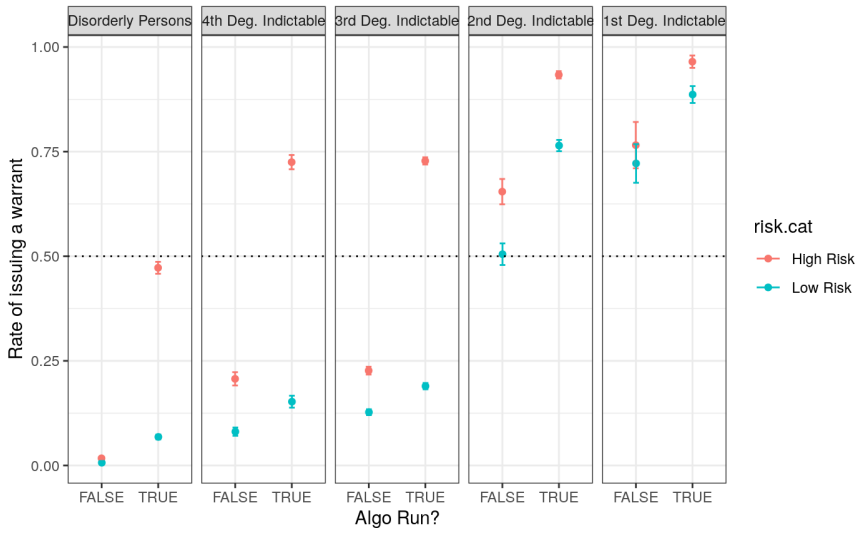

Figure 4.9. Warrants, charges, and algorithmic risk This graph shows how warrant rates differ by the algorithmic risk score of the defendant, the defendant's current charge, and whether the officer viewed the risk scores. Each panel corresponds to a different classification of crime—classifications are ordered from left to right by increasing severity. Within each panel, dots in the left column indicate warrant rates when the officer has not seen the scores; and the right columns indicate warrant rates when the officer has seen the scores. Defendants are labeled "High Risk" if either score (failure to appear or new criminal activity) is above the algorithmic recommendation threshold of 3 out of 6. Defendants are labeled "Low Risk" if both scores are below the threshold.

Figure 4.9 starts to suggest the room for improvement in social efficiency. Many high risk defendants, low crime defendants are released, especially if the algorithm isn't run while many low risk, first degree indictable defendants are detained even when the algorithm **is** run.

4.5 Results

One central message from Section 4.4 is that charge severity matters for officer behavior and this preference seems inefficient. However, there are two weaknesses in such an argument. First, while descriptive statistics can be helpful for guiding a high-level understanding of the setting, they often lack a clear causal interpretation without including the right set of controls. Second, it is difficult to make normative judgments on the private preferences of

the officers, an insight on omitted payoff bias discussed recently by Kleinberg et al. (2015), Kleinberg et al. (2018b), and Rambachan (2021).

To address both of these concerns I focus on a factor with more straightforward normative implications: race. Adding formality to the graphs from Section 4.4, I run a battery of regression-based tests for discrimination in information acquisition, warrants, and belief updating.

4.5.1 Does race affect information acquisition?

Table 4.2 formally tests for discrimination in information acquisition in a regression framework by controlling for increasingly strict covariates. First, in Column 1 I document a raw racial disparity. White and female defendants are less likely to have a PSA run for their case. This gap could be explained by myriad reasons other than discrimination thus the succeeding columns control for various plausibly risk-related attributes of the defendant. When controlling for the details of the crime, the estimated gap falls to X; when controls for the risk score are added the race gap is Y; when controls for year, month, and officer are added, the gap persists at Z%. The last column of Table X, my preferred specification, indicates that, within officer, the attention discrimination that minority defendants face is X percentage points.

4.5.2 What effect does attention discrimination have on defendants?

Based on these results alone it is unclear whether the documented discrimination helps or harms the defendants. To address this uncertainty, Table 4.3 documents a parallel analysis where I replace the dependent variable from table 1 with a binary variable indicating whether the officer issued a warrant. Again, the first column indicates a racial disparity in the raw warrant rate percentages and the final column estimates differential treatment by a single officer after controlling for year, month, risk, and charge characteristics. While informative,

Table 4.3 likely attenuates the parameters of interests since it includes all arrests, including the situations where the officer has chosen not to see the risk score.

To address this shortcoming, I conduct a third set of analyses in Table 4.4 replicating Table 4.3 only for the cases where the PSA was run. Columns 1-4 in Table 4.4 are qualitatively identical to those in Table 4.3. Officers are more likely to issue warrants for racial minorities after controlling for all observables. To test for differential belief updating, I add an additional column, column 5, interacting the race indicator with each of the predicted risk scores. I find that officers are less sensitive to the risk scores for minority defendants, which results in a large warrant gap for the least risky defendants; the gap closes for riskier defendants. That the gap is largest for low-risk defendants has two direct policy implications. First, since the plurality of defendants are of the lowest risk, the largest gap affects the largest percentage of the minority defendants population. Second, the gap is nearly three times as large as the marginal effect of each risk score; the lowest-risk (risk score = 1) minority defendant recommended for release is treated approximately the same as a white defendant with a risk score of 4 who is recommended for detention.

4.6 Conclusion

A natural implication of the fact that attention is scare is that people allocate their attention strategically to support their goals and objectives. I provide empirical evidence that officers intentionally allocate their attention to further their interests, which seem to be broader than minimizing public safety risk.

My findings present several lessons for the large scale deployment of information, especially algorithmic decision aids. First, when the goals of the user are different from those of the algorithm designer or social planner, then **user error and manipulation** can lead to social welfare losses from misranking and other allocation mistakes. Note that this result does not depend on any notion of algorithmic bias, neither demographic nor statistical. This implies

that strategic users can render even an optimized algorithm with no societal biases as tools for social inefficiency and discrimination.

Second, although this paper replicates the finding that an automated algorithmic system can outperform a human, I emphasize that this paper does not and cannot reject efficient human + algorithm will outperform full automation. Beyond the likely political resistance to full automation, there is good reason not to prefer automation. In short, in most settings, there are likely potential information gains to be had from **private information**. The human will almost necessarily "see" something that the algorithm *could not*. Note that this is a more restrictive condition than the human using information that the algorithm *does not* use. When the algorithm was trained it, in principle, could have had access to factors such as race, gender, and the current charge. However these factors are not used in the algorithm either because the designer deemed those categories socially unacceptable or the algorithm deemed them statistically unuseful. For these reasons, allowing the officers to make unstructured deviations based on such factors is unlikely to increase social welfare. On the other hand, the algorithm could never know whether the defendant threatens to harm their partner if released and such signals that are unobservable by the algorithm should be incorporated.

Finally, having a structure and guidance about disciplined **behavioral override** would improve human+machine decision making at large. What that means for New Jersey and similar jurisdictions is to have more friction and accountability around overriding the algorithm. Only mitigating private information should be grounds for overriding the algorithmic recommendation and such reasoning should be justified and audited with regularity. For example, New Jersey could mandate that officers consult the algorithm and follow what it says unless they can document a specific reason that the recommendation should be overridden. Then on a regular basis those exemptions can be assessed for how predictive they are of risk.

Beyond the criminal justice space, algorithms surround us increasingly and influence disparate choices from how we shop online to how we drive our cars. Knowing when to ask for directions and then knowing when to do what we are told will be key to getting the most from our algorithms, without getting lost.

4.7 Tables

Table 4.1
Descriptive Summary Statistics by Information Acquisition

This table summarizes basic details about the demographics, arrest details, risk level and warrant decisions for arrests in New Jersey from 2018 to 2021. The sample is split by whether the Public Safety Assessment was run or not.

		FALSE (N=61293)		TRUE (N=46113)	
		Mean	Std. Dev.	Mean	Std. Dev.
Max PSA Score		3.2	1.7	3.3	1.7
Age at arrest		33.0	11.8	33.2	11.2
		N	Pct.	N	Pct.
Warrant issued?	FALSE	55486	90.5	24482	53.1
	TRUE	5807	9.5	21631	46.9
PSA recommendation	Release recommended	36082	58.9	24834	53.9
	Warrant recommended	25211	41.1	21279	46.1
Charge severity	Disorderly Persons	36315	59.2	11999	26.0
	4th Deg. Indictable	5417	8.8	5079	11.0
	3rd Deg. Indictable	16578	27.0	20578	44.6
	2nd Deg. Indictable	2394	3.9	6907	15.0
	1st Deg. Indictable	589	1.0	1550	3.4
Charge type	drugs	34816	56.8	27222	59.0
	nan	217	0.4	3	0.0
	other	4709	7.7	2017	4.4
	person	3746	6.1	3121	6.8
	property	16085	26.2	10277	22.3
	public_order	724	1.2	288	0.6
	weapons	996	1.6	3185	6.9
Gender	Male	45388	74.1	35981	78.0
	Female	15903	25.9	10128	22.0
Race	White	8302	13.5	8813	19.1
	Black	10640	17.4	13635	29.6
	Latino	4009	6.5	5298	11.5
	Asian	125	0.2	196	0.4
	Native American	6	0.0	12	0.0
	Other	35	0.1	39	0.1
	Unknown	38176	62.3	18120	39.3

Table 4.2
Information Acquisition Regression Results

	Dependent variable:			
	ran_PSA			
	OLS			*felm*
	(1)	(2)	(3)	(4)
	base	charges	risk	kitchen.sink
simple_raceWhite	−0.047***	−0.028***	−0.027***	−0.023***
	(0.005)	(0.004)	(0.004)	(0.004)
diag_genderFemale	−0.018***	−0.003	−0.004	−0.014***
	(0.003)	(0.003)	(0.003)	(0.003)
charge_type4th Deg. Indictable		0.261***	0.262***	0.238***
		(0.005)	(0.005)	(0.005)
charge_type3rd Deg. Indictable		0.278***	0.279***	0.282***
		(0.003)	(0.003)	(0.003)
charge_type2nd Deg. Indictable		0.404***	0.394***	0.404***
		(0.006)	(0.006)	(0.006)
charge_type1st Deg. Indictable		0.394***	0.379***	0.445***
		(0.010)	(0.010)	(0.010)
clny_derived_psa__nca_score			−0.004**	−0.001
			(0.002)	(0.002)
clny_derived_psa__fta_score			−0.013***	−0.003*
			(0.002)	(0.001)
Constant	0.567***	0.394***	0.457***	
	(0.003)	(0.003)	(0.005)	
Observations	107,406	107,406	107,406	107,406
R^2	0.053	0.156	0.158	0.555
Adjusted R^2	0.053	0.156	0.158	0.489
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4.3
Warrant Regression Results - Full Sample

	Dependent variable:			
	warrant_issued			
	OLS			*felm*
	(1)	(2)	(3)	(4)
	base	charges	risk	kitchen.sink
simple_raceWhite	−0.082***	−0.029***	−0.028***	−0.031***
	(0.003)	(0.003)	(0.003)	(0.003)
diag_genderFemale	−0.032***	−0.026***	−0.022***	−0.018***
	(0.003)	(0.002)	(0.002)	(0.002)
charge_type4th Deg. Indictable		0.126***	0.125***	0.110***
		(0.004)	(0.004)	(0.004)
charge_type3rd Deg. Indictable		0.179***	0.179***	0.156***
		(0.002)	(0.002)	(0.002)
charge_type2nd Deg. Indictable		0.472***	0.481***	0.437***
		(0.004)	(0.004)	(0.005)
charge_type1st Deg. Indictable		0.591***	0.605***	0.507***
		(0.007)	(0.007)	(0.008)
clny_derived_psa__nca_score			0.016***	0.016***
			(0.001)	(0.001)
clny_derived_psa__fta_score			0.001	0.011***
			(0.001)	(0.001)
Constant	0.539***	0.314***	0.247***	
	(0.002)	(0.002)	(0.004)	
Observations	107,406	107,406	107,406	107,406
R^2	0.303	0.436	0.439	0.598
Adjusted R^2	0.303	0.436	0.439	0.539
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4.4
Warrant regression results - Conditional on PSA run

	warrant_issued				
	(1)	(2)	(3)	(4)	(5)
White	−0.075***	−0.020***	−0.022***	−0.040***	−0.095***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.013)
Female	−0.025***	−0.019***	−0.011***	−0.005	−0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
New Crime Risk (1-6)			0.027***	0.031***	0.039***
			(0.002)	(0.002)	(0.003)
Failure to Appear Risk (1-6)			0.027***	0.035***	0.038***
			(0.002)	(0.002)	(0.003)
White * New Crime Risk (1-6)					−0.005
					(0.005)
White * Failure to Appear Risk (1-6)					0.020***
					(0.005)
Constant	0.771***	0.585***	0.368***		
	(0.003)	(0.004)	(0.006)		
N	46,113	46,113	46,113	46,113	46,113
Adjusted R^2	0.468	0.519	0.540	0.604	0.608

Notes:
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 4.5
Warrant regression results - Conditional on PSA not run

	warrant_issued				
	(1)	(2)	(3)	(4)	(5)
White	−0.038***	−0.020***	−0.019***	−0.010***	0.001
	(0.004)	(0.003)	(0.003)	(0.003)	(0.009)
Female	−0.023***	−0.028***	−0.028***	−0.018***	−0.018***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
New Crime Risk (1-6)			0.002	0.001	−0.005**
			(0.001)	(0.001)	(0.002)
Failure to Appear Risk (1-6)			−0.011***	−0.003**	−0.006***
			(0.001)	(0.001)	(0.002)
White * New Crime Risk (1-6)					−0.002
					(0.004)
White * Failure to Appear Risk (1-6)					−0.001
					(0.003)
Constant	0.236***	0.090***	0.123***		
	(0.002)	(0.002)	(0.004)		
N	61,291	61,291	61,291	61,291	61,291
Adjusted R^2	0.110	0.312	0.315	0.563	0.564

Notes:
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

REFERENCES

Agan, Amanda and Sonja Starr, 2018, Ban the box, criminal records, and racial discrimination: A field experiment, *The Quarterly Journal of Economics* 133, 191–235.

Aigner, Dennis J and Glen G Cain, 1977, Statistical theories of discrimination in labor markets, *Ilr Review* 30, 175–187.

Allcott, Hunt, 2011, Social norms and energy conservation, *Journal of public Economics* 95, 1082–1095.

Allen, Eric J, Patricia M Dechow, Devin G Pope, and George Wu, 2017, Reference-dependent preferences: Evidence from marathon runners, *Management Science* 63, 1657–1672.

Alvídrez, Salvador, Valeriano Piñeiro-Naval, María Marcos-Ramos, and José Luis Rojas-Solís, 2015, Intergroup contact in computer-mediated communication: The interplay of a stereotype-disconfirming behavior and a lasting group identity on reducing prejudiced perceptions, *Computers in Human Behavior* 52, 533–540.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, 2016, Machine bias, in *Ethics of Data and Analytics*, 254–264 (Auerbach Publications).

Anik, Lalin and Michael I Norton, 2020, On being the tipping point: Social threshold incentives motivate behavior, *Journal of the Association for Consumer Research* 5, 19–33.

Antheunis, Marjolijn L, Mariek MP Vanden Abeele, and Saskia Kanters, 2015, The impact of facebook use on micro-level social capital: A synthesis, *Societies* 5, 399–419.

Arkes, Hal R and Philip E Tetlock, 2004, Attributions of implicit prejudice, or" would jesse jackson'fail'the implicit association test?", *Psychological inquiry* 15, 257–278.

Arrow, Kenneth J., 1973, The theory of discrimination, in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, 3–33.

Asch, Solomon E, 1955, Opinions and social pressure, *Scientific American* 193, 31–35.

Baer, Markus, 2010, The strength-of-weak-ties perspective on creativity: a comprehensive examination and extension., *Journal of applied psychology* 95, 592.

Bang, Dan and Chris D Frith, 2017, Making better decisions in groups, *Royal Society open science* 4, 170193.

Bansak, Kirk, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein, 2018, Improving refugee integration through data-driven algorithmic assignment, *Science* 359, 325–329.

Bargh, John A, 1994, The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition, *Handbook of social cognition* 1, 1–40.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka, 2016, Attention discrimination: Theory and field experiments with monitoring information acquisition, *American Economic Review* 106, 1437–75.

Becker, Gary, 1957, *The Economics of Discrimination* (University of Chicago Press).

Bento, Francisco Ramadas da Silva Ribeiro, 2018, *Predicting start-up success with machine learning*, Ph.D. thesis.

Benz, Matthias and Stephan Meier, 2008, Do people behave in experiments as in the field?—evidence from donations, *Experimental economics* 11, 268–281.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan, 2005, Implicit discrimination, *American Economic Review* 95, 94–98.

Bertrand, Marianne and Esther Duflo, 2017, Field experiments on discrimination, *Handbook of economic field experiments* 1, 309–393.

Bertrand, Marianne and Sendhil Mullainathan, 2004, Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination, *American Economic Review* 94, 991–1013.

Blackwell, David, 1953, Equivalent comparisons of experiments, *The annals of mathematical statistics* 265–272.

Bloom, Nicholas and John Van Reenen, 2007, Measuring and explaining management practices across firms and countries, *The quarterly journal of Economics* 122, 1351–1408.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope, 2019, Inaccurate statistical discrimination: An identification problem, Tech. rep., National Bureau of Economic Research.

Buolamwini, Joy and Timnit Gebru, 2018, Gender shades: Intersectional accuracy disparities in commercial gender classification, in *Conference on fairness, accountability and transparency*, 77–91 (PMLR).

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott, 2020, Misperceived social norms: Women working outside the home in saudi arabia, *American economic review* 110, 2997–3029.

Camerer, Colin F, 2011, *Behavioral game theory: Experiments in strategic interaction* (Princeton university press).

Cameron, Kim and Jane Dutton, 2003, *Positive organizational scholarship: Foundations of a new discipline* (Berrett-Koehler Publishers).

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez, 2012, Inequality at work: The effect of peer salaries on job satisfaction, *American Economic Review* 102, 2981–3003.

Carmeli, Abraham, Batia Ben-Hador, David A Waldman, and Deborah E Rupp, 2009, How leaders cultivate social capital and nurture employee vigor: Implications for job performance., *Journal of Applied Psychology* 94, 1553.

Carmeli, Abraham, Daphna Brueller, and Jane E Dutton, 2009, Learning behaviours in the workplace: The role of high-quality interpersonal relationships and psychological safety, *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research* 26, 81–98.

Cialdini, Robert B and Noah J Goldstein, 2004, Social influence: Compliance and conformity, *Annual review of psychology* 55, 591–621.

Collins, Jim, 2009, Good to great-(why some companies make the leap and others don't).

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, 2017, Algorithmic decision making and the cost of fairness, in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Cunningham, Thomas, 2013, Biases and implicit knowledge .

Dawes, Robyn M, David Faust, and Paul E Meehl, 1989, Clinical versus actuarial judgment, *Science* 243, 1668–1674.

DellaVigna, Stefano and Matthew Gentzkow, 2019, Uniform pricing in us retail chains, *The Quarterly Journal of Economics* 134, 2011–2084.

Deutsch, Morton and Harold B Gerard, 1955, A study of normative and informational social influences upon individual judgment., *The journal of abnormal and social psychology* 51, 629.

Dhar, Ravi, 1997, Consumer preference for a no-choice option, *Journal of consumer research* 24, 215–231.

Di Maggio, Marco, Dimuthu Ratnadiwakara, and Don Carmichael, 2022, Invisible primes: Fintech lending with alternative data, Tech. rep., National Bureau of Economic Research.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, 2015, Algorithm aversion: people erroneously avoid algorithms after seeing them err., *Journal of Experimental Psychology: General* 144, 114.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, 2018, Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them, *Management Science* 64, 1155–1170.

Ellison, Nicole B, Charles Steinfield, and Cliff Lampe, 2007, The benefits of facebook "friends:" social capital and college students' use of online social network sites, *Journal of computer-mediated communication* 12, 1143–1168.

Enke, Benjamin, 2020, Wysiati, *Quarterly Journal of Economics* .

Erel, Isil, Léa H Stern, Chenhao Tan, and Michael S Weisbach, 2021, Selecting directors using machine learning, *The Review of Financial Studies* 34, 3226–3264.

Fama, Eugene F, 1965, The behavior of stock-market prices, *The journal of Business* 38, 34–105.

Fitzsimons, Gráinne M, Esther Sackett, and Eli J Finkel, 2016, Transactive goal dynamics theory: A relational goals perspective on work teams and leadership, *Research in Organizational Behavior* 36, 135–155.

Fudenberg, Drew and David K Levine, 2006, A dual-self model of impulse control, *American economic review* 96, 1449–1476.

Gabaix, Xavier, 2019, Behavioral inattention, in *Handbook of Behavioral Economics: Applications and Foundations 1*, vol. 2, 261–343 (Elsevier).

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–74.

Gentzkow, Matthew and Jesse M Shapiro, 2010, What drives media slant? evidence from us daily newspapers, *Econometrica* 78, 35–71.

Ghassemi, M, C Song, and T Alhanai, 2020, The automated venture capitalist: Data and methods to predict the fate of startup ventures., in *AAAI KDF Workshop*.

Golman, Russell, David Hagmann, and George Loewenstein, 2017, Information avoidance, *Journal of Economic Literature* 55, 96–135.

Gompers, Paul, Steven N Kaplan, and Vladimir Mukharlyamov, 2015, What do private equity firms say they do?, Tech. rep., National Bureau of Economic Research.

Gompers, Paul A, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev, 2020, How do venture capitalists make decisions?, *Journal of Financial Economics* 135, 169–190.

Granovetter, Mark S, 1973, The strength of weak ties, *American journal of sociology* 78, 1360–1380.

Greenwald, Anthony G, Debbie E McGhee, and Jordan LK Schwartz, 1998, Measuring individual differences in implicit cognition: the implicit association test., *Journal of personality and social psychology* 74, 1464.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, 2014, Learning through noticing: Theory and evidence from a field experiment, *The Quarterly Journal of Economics* 129, 1311–1353.

Hannaford-Agor, Paula L, Valerie P Hans, Nicole L Mott, and G Thomas Munsterman, 2002, Are hung juries a problem, *Williamsburg, VA: National Center for State Courts* .

Heaton, Alan W and Arie W Kruglanski, 1991, Person perception by introverts and extraverts under time pressure: Effects of need for closure, *Personality and Social Psychology Bulletin* 17, 161–165.

Hehman, Eric, Jessica K Flake, and Jimmy Calanchini, 2018, Disproportionate use of lethal force in policing is associated with regional racial biases of residents, *Social psychological and personality science* 9, 393–401.

Hellmann, Thomas and Manju Puri, 2000, The interaction between product market and financing strategy: The role of venture capital, *The review of financial studies* 13, 959–984.

Hellmann, Thomas and Manju Puri, 2002, Venture capital and the professionalization of start-up firms: Empirical evidence, *The journal of finance* 57, 169–197.

Hogarth, Robin M, Tomás Lejarraga, and Emre Soyer, 2015, The two settings of kind and wicked learning environments, *Current Directions in Psychological Science* 24, 379–385.

Hortaçsu, Ali and Steven L Puller, 2008, Understanding strategic bidding in multi-unit auctions: a case study of the texas electricity spot market, *The RAND Journal of Economics* 39, 86–114.

Howell, Sabrina T, 2020, Reducing information frictions in venture capital: The role of new venture competitions, *Journal of Financial Economics* 136, 676–694.

Hu, Allen and Song Ma, 2021, Persuading investors: A video-based study, Tech. rep., National Bureau of Economic Research.

Isenberg, Daniel J, 1986, Group polarization: A critical review and meta-analysis., *Journal of personality and social psychology* 50, 1141.

Jamieson, David W and Mark P Zanna, 1989, Need for structure in attitude formation and expression, *Attitude structure and function* 383–406.

Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon, 2016, Combining satellite imagery and machine learning to predict poverty, *Science* 353, 790–794.

Kahneman, Daniel, 2011, *Thinking, fast and slow* (Macmillan).

Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein, 2021, *Noise: A flaw in human judgment* (Little, Brown).

Kaplan, Steven N, Mark M Klebanov, and Morten Sorensen, 2012, Which ceo characteristics and abilities matter?, *The Journal of Finance* 67, 973–1007.

Kaplan, Steven N and Antoinette Schoar, 2005, Private equity performance: Returns, persistence, and capital flows, *The journal of finance* 60, 1791–1823.

Kaplan, Steven N, Berk A Sensoy, and Per Strömberg, 2009, Should investors bet on the jockey or the horse? evidence from the evolution of firms from early business plans to public companies, *The Journal of Finance* 64, 75–115.

Kaplan, Steven N and Morten Sorensen, 2017, Are ceos different? characteristics of top managers, Tech. rep., National Bureau of Economic Research.

Kaplan, Steven N and Per Stromberg, 2001, Venture capitals as principals: Contracting, screening, and monitoring, *American Economic Review* 91, 426–430.

Kaplan, Steven N and Per ER Strömberg, 2004, Characteristics, contracts, and actions: Evidence from venture capitalist analyses, *The Journal of Finance* 59, 2177–2210.

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman, 2016, Getting to the top of mind: How reminders increase saving, *Management Science* 62, 3393–3411.

Kawachi, Ichiro, Lisa Berkman, et al., 2000, Social cohesion, social capital, and health, *Social epidemiology* 174, 290–319.

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu, 2019, Predicting returns with text data, Tech. rep., National Bureau of Economic Research.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, 2018a, Human decisions and machine predictions, *The Quarterly Journal of Economics* 133, 237–293.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, 2018b, Human decisions and machine predictions, *The quarterly journal of economics* 133, 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, 2015, Prediction policy problems, *American Economic Review* 105, 491–95.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan, 2022, The challenge of understanding what users want: Inconsistent preferences and engagement optimization, *arXiv preprint arXiv:2202.11776* .

Kogan, Shimon, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith, 2009, Predicting risk from financial reports with regression, in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–280.

Krishna, Amar, Ankit Agrawal, and Alok Choudhary, 2016, Predicting the outcome of startups: less failure, more success, in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 798–805 (IEEE).

Kruglanski, Arie W and Tallie Freund, 1983, The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring, *Journal of experimental social psychology* 19, 448–468.

Kruglanski, Arie W, Antonio Pierro, Lucia Mannetti, and Eraldo De Grada, 2006, Groups as epistemic providers: need for closure and the unfolding of group-centrism., *Psychological review* 113, 84.

Kruglanski, Arie W and Donna M Webster, 2018, Motivated closing of the mind:"seizing" and "freezing", *The motivated mind* 60–103.

Kunda, Ziva, 1990, The case for motivated reasoning., *Psychological bulletin* 108, 480.

Lerner, Josh, 1995, Venture capitalists and the oversight of private firms, *the Journal of Finance* 50, 301–318.

Lerner, Joshua, 1994, The syndication of venture capital investments, *Financial management* 16–27.

Levy, Leo, 1960, Studies in conformity behavior: a methodological note, *The Journal of Psychology* 50, 39–41.

List, John A, 2003, Does market experience eliminate market anomalies?, *The Quarterly Journal of Economics* 118, 41–71.

List, John A, 2006, The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions, *Journal of political Economy* 114, 1–37.

Logg, Jennifer M, Julia A Minson, and Don A Moore, 2019, Algorithm appreciation: People prefer algorithmic to human judgment, *Organizational Behavior and Human Decision Processes* 151, 90–103.

Lord, Charles G, Lee Ross, and Mark R Lepper, 1979, Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence., *Journal of personality and social psychology* 37, 2098.

Luce, Mary Frances, 1998, Choosing to avoid: Coping with negatively emotion-laden consumer decisions, *Journal of consumer research* 24, 409–433.

Lueke, Adam and Bryan Gibson, 2015, Mindfulness meditation reduces implicit age and race bias: The role of reduced automaticity of responding, *Social Psychological and Personality Science* 6, 284–291.

Lyonnet, Victor and Léa H Stern, 2022, Venture capital (mis) allocation in the age of ai, *Fisher College of Business Working Paper* 002.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, 2011, Learning word vectors for sentiment analysis, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150 (Association for Computational Linguistics, Portland, Oregon, USA).

Meehl, Paul E, 1954, Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. .

Milkman, Katherine L, Modupe Akinola, and Dolly Chugh, 2012, Temporal distance and discrimination: An audit study in academia, *Psychological science* 23, 710–717.

Moore, Don A and Paul J Healy, 2008, The trouble with overconfidence., *Psychological review* 115, 502.

Mullainathan, Sendhil and Ziad Obermeyer, 2019, A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions, *NBER Working Paper* .

Mullainathan, Sendhil and Ziad Obermeyer, 2022, Diagnosing physician error: A machine learning approach to low-value health care, *The Quarterly Journal of Economics* 137, 679–727.

Nanda, Ramana, Sampsa Samila, and Olav Sorenson, 2020, The persistent effect of initial success: Evidence from venture capital, *Journal of Financial Economics* 137, 231–248.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, 2019, Dissecting racial bias in an algorithm used to manage the health of populations, *Science* 366, 447–453.

Orchard, Jacob and Joseph Price, 2017, County-level racial prejudice and the black-white gap in infant health outcomes, *Social science & medicine* 181, 191–198.

Patacchini, Eleonora and Yves Zenou, 2008, The strength of weak ties in crime, *European Economic Review* 52, 209–236.

Payne, B Keith, Alan J Lambert, and Larry L Jacoby, 2002, Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons, *Journal of Experimental Social Psychology* 38, 384–396.

Penrod, Steven D, Nancy Pennington, and Reid Hastie, 1983, Inside the jury.

Perez-Truglia, Ricardo and Ugo Troiano, 2018, Shaming tax delinquents, *Journal of Public Economics* 167, 120–137.

Phelps, Edmund S., 1972, The statistical theory of racism and sexism, *American Economic Review* 62, 659–61.

Putnam, Robert D, 2000, Bowling alone: America's declining social capital, in *Culture and politics*, 223–234 (Springer).

Rajan, Raghuram G, 2012, Presidential address: The corporation in finance, *The Journal of Finance* 67, 1173–1217.

Rambachan, Ashesh, 2021, Identifying prediction mistakes in observational data.

Rees-Jones, Alex, 2018, Quantifying loss-averse tax manipulation, *The Review of Economic Studies* 85, 1251–1278.

Sandstrom, Gillian M and Elizabeth W Dunn, 2014, Social interactions and well-being: The surprising power of weak ties, *Personality and Social Psychology Bulletin* 40, 910–922.

Sandys, Marla and C Dillehay, 1995, First-ballot votes, predeliberation dispositions, and final verdicts in jury trials, *Law and Human Behavior* 19, 175–195.

Schwartzstein, Joshua and Adi Sunderam, 2021, Using models to persuade, *American Economic Review* 111, 276–323.

Simon, Herbert A, 1955, A behavioral model of rational choice, *The quarterly journal of economics* 69, 99–118.

Smith, Vicki L and Saul M Kassin, 1993, Effects of the dynamite charge on the deliberations of deadlocked mock juries, *Law and Human Behavior* 17, 625–643.

Spranca, Mark, Elisa Minsk, and Jonathan Baron, 1991, Omission and commission in judgment and choice, *Journal of experimental social psychology* 27, 76–105.

Strulov-Shlain, Avner, 2021, More than a penny's worth: Left-digit bias and firm pricing, *Chicago Booth Research Paper* .

Tassier, Troy, 2006, Labor market implications of weak ties, *Southern Economic Journal* 704–719.

Thaler, Richard H, 2018, Nudge, not sludge.

Tuncel, Ece, Alexandra Mislin, Selin Kesebir, and Robin L Pinkley, 2016, Agreement attraction and impasse aversion: Reasons for selecting a poor deal over no deal at all, *Psychological science* 27, 312–321.

Tversky, Amos and Eldar Shafir, 1992, Choice under conflict: The dynamics of deferred decision, *Psychological science* 3, 358–361.

Webster, Donna M and Arie W Kruglanski, 1994, Individual differences in need for cognitive closure., *Journal of personality and social psychology* 67, 1049.

Yeomans, Michael, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg, 2019, Making sense of recommendations, *Journal of Behavioral Decision Making* 32, 403–414.

Ying, Dafei, Gaosheng Wua, and Jing Lia, 2021, Comparison of different machine learning methods for predicting the success of the startups, Tech. rep., Sinovation Ventures.

Żbikowski, Kamil and Piotr Antosiuk, 2021, A machine learning, bias-free approach for predicting business success using crunchbase data, *Information Processing & Management* 58, 102555.

Zhang, Ye, 2020, Discrimination in the venture capital industry: Evidence from two randomized controlled trials, *arXiv preprint arXiv:2010.16084* .

Zingales, Luigi, 2000, In search of new foundations, *The journal of Finance* 55, 1623–1653.