Essays on Online Platforms and Human-Algorithm Interaction

by

Alex Moehring

B.S., University of North Carolina at Chapel Hill (2014)
S.M., Massachusetts Institute of Technology (2022)

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

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Alex Moehring Department of Management May 2, 2024
Certified by:	Catherine Tucker Department of Management Thesis Supervisor
Accepted by:	Eric So Professor, Global Economics and Finance Faculty Chair, MIT Sloan PhD Program

Essays on Online Platforms and Human-Algorithm Interaction

by

Alex Moehring

Submitted to the Department of Management
on May 2, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT

ABSTRACT

This dissertation contains three chapters that analyze how algorithms on social media platforms influence the content that users engage with and how individuals incorporate algorithmic predictions in their decision-making. In Chapter 1, I study how engagement maximizing news feed algorithms on social media affect the credibility of news content with which users engage. This allows me to estimate the extent to which engagement-maximizing algorithms promote and incentivize low-quality content. In addition, I evaluate how the ranking algorithm itself can be designed to promote and encourage engagement with high quality content. In Chapter 2, I analyze how the introduction of a new non-personalized news feed impacts user engagement quantity, quality, and diversity on the Reddit platform. I find that this auxiliary feed increases the share of users that engage with news-related content and the diversity of engagement within news categories and within articles from publishers across the political spectrum increases as a result of the feed. In Chapter 3, in collaboration with Nikhil Agarwal, Tobias Salz, and Pranav Rajpurkar, we study human-AI collaboration using an information experiment with professional radiologists. Results show that providing (i) AI predictions does not always improve performance, whereas (ii) contextual information does. Radiologists do not realize the gains from AI assistance because of errors in belief updating – they underweight AI predictions and treat their own information and AI predictions as statistically independent.

Thesis supervisor: Catherine Tucker
Title: Sloan Distinguished Professor of Management

Acknowledgments

I have been extremely fortunate to study at MIT and my growth as a researcher would not be possible without the wonderful community at MIT. First, I want to thank my advisor, Catherine Tucker. She has encouraged me to pursue ambitious projects and showed me how to do so in a rigorous manner. She has also looked out for me at every step of the process and I am immensely grateful. I also thank my committee members Nikhil Agarwal and Dean Eckles for their guidance and advice. I learned so much working with them as a student, research assistant, and co-author and it has been a privilege to do so.

I also want to thank other faculty who have advised me along the way, including Sinan Aral, Erik Brynjolfsson, and Tobias Salz. In addition, I am grateful for all the faculty at MIT from whom I have learned so much. The faculty in MIT Sloan's Information Technology, Marketing, and Applied Economics groups and the MIT Economics department have all been particularly instrumental in my development.

I have been fortunate to work with numerous coauthors and learn from other PhD students. I am thankful in particular to the IT students including Daniel Rock, Avinash Collis, David Holtz, Michael Zhao, Sebastian Steffen, Zanele Munyikwa, Emma Wiles, Hong-Yi Tu Ye, Jaeyoon Song, Hirotaka Miura, Mohammad Alsobay, Michael Caosun, Justin Kaashoek, Ben Manning, Robin Na, Anand Shah, and Peyman Shahidi. In addition, I am grateful for Tim Di Silva, Luca Gius, and Carlos Molina who have been both great friends and colleagues and a great help in teaching me economics during our first years.

I am also thankful to the Initiative on the Digital Economy staff and post-docs who I was fortunate to work with including Ananya Sen, Paramveer Dhillon, Sarah Bana, Georgios Petropoulos, Xiupeng Wang, Seth Benzell, Madhav Kumar, Harang Ju, and Ehsan Valavi. I also thank the doctoral program staff: Hillary Ross, Davin Schnappauf, and Lauren Koduah.

Before MIT, I had the good fortune of being advised by many amazing mentors and teachers. Specifically, I would like to thank Mike Aguilar, Markus Mobius, and Greg Lewis.

Finally, I thank my friends and family for their support throughout my life and my graduate studies. In particular, I thank my parents Richard and Carol Anne Moehring, my sister and brother-in-law Jenna Moehring and Damon D, and my friends from home. Most importantly, I would like to thank my wife MacKenzie Walker for her love and support during my PhD and beyond. Without you this dissertation would not be possible.

Contents

[1] This chapter is written in collaboration with Nikhil Agarwal, Tobias Salz, and Pranav Rajpurkar [Agarwal et al., 2023].

List of Figures

List of Tables

Chapter 1

Personalized Rankings and User Engagement: An Empirical Evaluation of the Reddit News Feed

1.1 Introduction

Digital platforms curate content for their users because of limited user attention and the vast amount of available content. The advertising business model adopted by many platforms creates an incentive to promote content through ranking algorithms that predict what content users are most likely to act on via clicking, liking, or commenting [Narayanan, 2023, Thorburn et al., 2022]. Personalized ranking algorithms that optimize for such engagement metrics may also promote low-quality or problematic content [Orlowski, 2020], which can negatively impact platforms if concerns over brand safety lead advertisers to respond by reducing advertising spending [Ahmad et al., 2023] or if there is a disconnect between short-term engagement metrics and long-term user welfare [Agan et al., 2023, Allcott et al., 2022, Kleinberg et al., 2022, Spence and Owen, 1977]. Moreover, concerns that ranking algorithms promote and incentivize low-quality content have prompted policy makers around the world to consider regulating ranking algorithms. Therefore, managers must balance maximizing engagement with the health of the platform ecosystem to satisfy both internal and external stakeholders.

There is an active debate surrounding the benefits and potential risks of personalized rankings that optimize for engagement. Platform managers often contend that ranking algorithms act as agents for users by promoting a user's preferred content and reducing search frictions on the platform [Dorsey, 2022]. Critics, however, frequently raise concerns that optimizing for engagement can incentivize low-quality content and reduce the diversity of viewpoints to which users are exposed [Orlowski, 2020, Pariser, 2011]. Despite these competing narratives, the impact of personalized news feeds on the quality of content users engage with remains an important and largely unresolved question. The lack of evidence regarding these issues primarily stems from the substantial challenges to studying ranking algorithms on social media platforms, including platforms' hesitance to share data and experiments with external researchers [Eckles, 2022].

This paper studies the impact personalized ranking algorithms that optimize for engagement have on the quality of content that is promoted to users and with which users engage. I explore this question in the context of political news on Reddit. In particular, I focus on the platform's largest politics community that centers on sharing and discussing news articles about US political news. In this community, users share news articles about US politics and then engage in discussion and commentary in comment threads alongside each article. I use the number of comments an article receives as the primary measure of engagement.

Reddit is among the ten most popular social media platforms in the world [SimilarWeb, 2024]. In addition, Reddit is an important component of the digital ecosystem playing an influential role in both web search and a source of training data for large language models [Patel, 2024]. The Reddit politics community I study also provides an ideal laboratory to analyze the types of content that are promoted under alternative ranking algorithms. The community is important to the platform due its size and the strong preferences of advertisers to not appear alongside low-quality news content [Ahmad et al., 2023].[1] In addition, it is often challenging to evaluate the quality of content on social media. Studying the politics community, which focuses on discussing news articles, allows me to use established measures of publisher credibility as a neutral measure of quality [Lin et al., 2022]. Many platform stakeholders, including advertisers, employees, and policymakers, have demonstrated an interest in the credibility of news content that appears on platforms which motivates the focus on publisher credibility in this analysis. Moreover, this community contains substantial heterogeneity in content quality – the credibility rating of the article's publisher – and horizontal differentiation in content based on the political slant of the article's publisher. There is also substantial heterogeneity in user preferences. Taken together, these sources of heterogeneity, which are common on many social media platforms, make it difficult to predict ex-ante what impact personalization and optimizing for engagement will have on the credibility of content that is promoted.

Throughout the analysis, a central challenge will be that a post's rank – its position in the feed – is endogenous. One should be concerned that a post's potential outcomes are correlated with its position in the feed, as I expect the existing feed to promote posts that are

[1] The politics community is consistently ranked as one of the most active communities on the platform.

more 'commentable' relative to posts that are not promoted. Therefore, to identify position effects – the causal effect a post's position has on the number of comments it receives – I exploit a novel regression discontinuity revealed in an open-source mirror of the platform's code base. This open-source mirror allows me to inspect the ranking algorithm and recreate the numerical score that is used to rank posts. Consequently, this permits using a regression discontinuity design to identify the local average treatment effect of a post's rank on the number of comments it receives in a period. As the ranking score of a focal post passes the score of a competing post, there is a discontinuous jump in the probability the focal post is ranked lower on the page.[2] The treatment effect estimates suggest that the causal effect of a post being promoted from the second position in the feed to the first position results in a 42.5% increase in the number of comments the post receives in a period. The effect of being promoted declines further down the feed, as the causal effect of moving one position higher on the feed is largest for the first position.

With an identification strategy for position effects, I turn to understanding the impact of optimizing for engagement with personalized rankings on the type of content to which users are exposed and with which they engage. To do so, I estimate a micro-founded model of user comment decisions. I model engagement decisions based on two components: whether a user is exposed to a post and whether, conditional on exposure, their utility from commenting exceeds the utility of the outside option. Post rank impacts engagement in this model only through the exposure component, where the probability of being exposed to a post depends on the post's rank. Conditional on exposure, users then have heterogeneous preferences to comment on posts depending on the political slant and credibility rating of the publisher. This model is identified using the reduced form position effect estimates and individual engagement choices. I use the model to estimate engagement patterns under counterfactual ranking algorithms including both personalized and non-personalized engagement maximization. In addition, I evaluate alternative credibility-aware ranking algorithms that optimizes for an objective function that balances total engagement and engagement with high-credibility publishers.

I find that a personalized engagement maximizing algorithm exacerbates differences in the share of user engagement with high-credibility publishers. An engagement maximizing algorithm tends to promote high-credibility publishers to users engaging with high-credibility publishers under Reddit's actual ranking algorithm and promotes lower-credibility publishers to users engaging with less-credible publishers under the actual ranking algorithm. This

[2] This identification strategy is most closely related to Narayanan and Kalyanam [2015], where data on the AdRank scores in Google auctions are used to estimate the position effects on Google advertisements, though to my knowledge this is the first application of such a strategy to a social media setting.

result indicates that personalization and engagement maximization increase heterogeneity in news diet quality – the share of a user's engagement with high-credibility publishers – and can lead to a subset of users being responsible for much of the engagement with low-credibility content on the platform. Moreover, I find that personalized engagement maximization leads to engagement with publishers that are less politically diverse and more similar to publishers the user has engaged with previously.

I also use the model of engagement decisions to analyze engagement patterns under counterfactual ranking algorithms that optimize alternative objective functions that explicitly trade-off total engagement and engagement with high-credibility publishers. At one extreme, this nests a credibility-maximizing algorithm that maximizes engagement with high-credibility publishers. This algorithm leads to a meaningful 5.0% decline in total user engagement. That said, platforms can achieve over half of the increase in news diet quality from the credibility-maximizing algorithm for a more modest 1.9% decrease in engagement. This change in engagement is similar in magnitude to the difference between the personalized and non-personalized engagement-maximizing algorithms. However, the non-personalized algorithm does not meaningfully improve the quality of user news diets, while the credibility-aware algorithm increases the average user's share of engagement with high-credibility publishers by 5.9 percentage points. This suggests there is room for managers to balance the competing objectives of maximizing total engagement and improving the health of the platform's ecosystem, and the ranking algorithm appears to be a useful tool to achieve such balance. Moreover, these findings highlight the potential benefits of personalization, as the personalized credibility-aware algorithm permits the platform to substantially increase the credibility of publishers promoted for the same quantity of engagement as the non-personalized engagement-maximizing algorithm.

An additional benefit of focusing on comments as the measure of engagement is the ability to analyze the text content to evaluate how comment sentiments change under counterfactual ranking algorithms. The results suggest that both the personalized and non-personalized engagement-maximizing algorithms slightly elevate the share of negative comments for the average user. Personalization, however, increases the dispersion of negative comment shares relative to the non-personalized algorithm. On average, users who prefer low-credibility publishers have the largest increases in their negative-sentiment engagement shares under the personalized algorithm. Given that negative comments contain strong negative emotions such as disgust and anger and are more likely to be classified as toxic, this finding suggests when low-credibility outlets drive user engagement it increases the likelihood of low-quality discussion.

Related Literature

This paper contributes primarily to two strands of the literature. A large literature has studied the impact of algorithmic recommendations on consumers. This literature considers algorithmic recommendations' impact on product sales [Donnelly et al., 2023, Fleder and Hosanagar, 2009, Ghose et al., 2014, Hosanagar et al., 2014, Lee and Hosanagar, 2019, Oestreicher-Singer and Sundararajan, 2012, Wang et al., 2023], content consumption [Aridor et al., 2022, Chen et al., 2023, Claussen et al., 2021, Holtz et al., 2020], the informational content of recommendations [Aridor et al., 2022], and consumer welfare [Chaney et al., 2018, Donnelly et al., 2023, Ghose et al., 2014]. In addition, this literature investigates how ranked feeds on social media platforms impact individual well-being [Kramer et al., 2014], media consumption [Bakshy et al., 2015a, Dujeancourt et al., 2021, Guess et al., 2023, Levy, 2021], and exposure to content from politicians [Huszár et al., 2022]. Much of this literature explores the impact of algorithmic ranking on the diversity of consumption [Berman and Katona, 2020, Chen et al., 2023, Claussen et al., 2021, Fleder and Hosanagar, 2009, Holtz et al., 2020, Van Alstyne and Brynjolfsson, 2005] or product sales [Hosanagar et al., 2014, Lee and Hosanagar, 2019, Oestreicher-Singer and Sundararajan, 2012] and how likely users are to be exposed to cross-cutting news publishers [Bakshy et al., 2015a, Levy, 2021]. Most related to this paper, Huszár et al. [2022] analyzes an experiment on Twitter and Guess et al. [2023] analyze an experiment on Facebook and Instagram that randomly assigns users to receive a reverse chronological ranking algorithm compared to those that received the existing personalized algorithm. Huszár et al. [2022] find that Twitter's personalized algorithm amplified right-leaning publishers. This is consistent with my finding that right leaning publishers see the largest increase in engagement in the personalized engagement-maximizing algorithm. Guess et al. [2023] find that the reverse chronological feed increases exposure to political and content from untrustworthy sources. While I find that a personalized engagement maximizing algorithm has little effect on average on the credibility of news content users are exposed to, I do find that for the majority of users engagement maximization slightly increases the share of engagement with high credibility sources. A contribution of this analysis is to unpack the heterogeneous effect of engagement maximizing algorithms on the credibility of content with which users engage. An additional contribution of the modeling-based approach taken here is that it allows me to evaluate alternative algorithms including a credibility-aware algorithm that balances optimizing for total engagement with engagement with high-credibility publishers.

Second, this paper contributes to the large and growing literature studying interventions to improve the quality of information people consume online. This literature both documents the reach of misinformation on social media and how it spreads [Allcott and Gentzkow, 2017, Grinberg et al., 2019, Guess et al., 2019, 2020, Vosoughi et al., 2018a] and evaluates interventions to curb the spread of misinformation (see Pennycook and Rand [2021] and Lazer et al. [2018] for a review). My findings are consistent with the literature showing that a minority of users account for a large share of consumption of low-credibility news content, and I contribute to the literature by finding that personalized engagement-maximizing algorithms exacerbate this difference [Allcott and Gentzkow, 2017, Grinberg et al., 2019, Guess et al., 2019, 2020]. In addition, this literature assesses many behavioral interventions through both lab and field experiments. That said, empirical evaluations of algorithmic interventions have been more difficult given limited access to platform data. I contribute to this literature by exploring how ranking algorithms affect the credibility of news with which users engage. More specifically, I study how personalization heterogeneously impacts the quality of users' news diets and consider the impact of algorithmic interventions on the credibility of news users engage with in addition to estimating the cost to the platform of foregone engagement of such interventions.

Implications for Managers and Policy Makers

These findings have important managerial implications. Given advertiser concerns over brand safety and a reluctance to appear alongside low-credibility news publishers [Ahmad et al., 2023], platform managers must balance total engagement with the credibility of content being promoted to users. Additional internal and external stakeholders, including policy makers and platform employees, have also demonstrated interest in reducing the spread of low-credibility content on digital platforms [Haugen, 2021, Warner, 2023]. The results presented here suggest codifying the trade-off explicitly in the objective function of the ranking algorithm is an effective method for limiting the spread of low-credibility news content. I estimate the cost of including credibility in the objective function and find that platforms willing to accept a modest decline in total engagement can substantially increase the share of engagement with high-quality publishers.

The findings also have important implications for policy makers. Concerns regarding ranking algorithms promoting and incentivizing low-quality content have prompted policy makers around the world to consider regulation that can address these issues. A common regulatory approach is to require platforms to allow users to opt out of personalized recommendations. The European Union's Digital Services Act includes such provisions, as do proposed laws in the United States such as the Filter Bubble Transparency Act. Related proposals in the United States that limit Section 230 protections for personalized recommendations would likely have a similar effect (e.g. Justice Against Malicious Algorithms

Act, and the Protecting Americans from Dangerous Algorithms Act). The implications of this study are clear: ranking algorithms can be designed to improve the credibility of the news content users engage with, and personalization is a valuable tool to mitigate the cost of moving away from optimizing only for engagement. A solution that takes advantage of the benefits of personalization, while protecting individual autonomy, would be to allow users to adjust the weights on different components within the ranking algorithm objective function, including the weight placed on publisher credibility. What weights users would choose and the results of such a design remain open questions and merit future work. Alternatively, regulators could incentivize platforms to align their ranking objective function with the preferences of society to take advantage of personalized ranking algorithms' substantial benefits while mitigating potential negative effects.

1.2 Background and Data

Reddit is a large social news aggregator with over 50 million daily active users as of January 2020 and was valued at approximately $6.5 billion in 2024 shortly after its initial public offering.[3] The platform is organized into over 100,000 virtual communities called subreddits that are focused on sharing and discussing content related to the community's topic. In this study, I focus on a subset of communities that are centered around sharing and discussing news articles. In these communities, users share news articles and then discuss the articles in comment threads as seen in Figure 1.1. Reddit is structured such that users can submit two types of content, submissions and comments. In the communities studied, submissions must contain a link to a news article and I therefore use the terms submissions, articles, and posts interchangeably. Users then discuss articles by posting comments – this commenting activity is the primary engagement measure I study.

1.2.1 Algorithmic Feeds on Reddit

Users interact with content on the platform via algorithmic feeds of a few different forms. Any user who visits a community page will see submissions from the community ranked by the platform's default ranking algorithm.[4] This algorithm sorts submissions according to the post's age and vote score – the net number of upvotes minus downvotes on a post – and is described in more detail in Section 1.3. In addition to the default algorithm, users can choose

[3] https://www.redditinc.com/

[4] On the platform, this default algorithm is called the *hot* algorithm. I refer to the hot algorithm as the actual ranking algorithm throughout.

to rank posts according to several alternative algorithms. The *new* algorithm implements a reverse chronological ranking; the *top* algorithm ranks posts according to the vote score in a given period; the *rising* algorithm favors recent posts; and the *controversial* algorithm promotes posts that have received more votes, either up or down, regardless of their direction. This paper focuses on the default algorithm's impact on engagement. All analyses presented here condition on the alternative algorithm rankings remaining unchanged. That is, when estimating the impact of post rank on engagement I estimate counterfactuals where post rank changes in the default feed but not in the alternative feeds.

In addition, Reddit users can join communities. Posts from these communities are displayed on a user's Home feed, the default feed users encounter when visiting the platform. The Home feed sorts posts according to the same default algorithm used by individual communities but ranks posts from all communities that a user is a member of rather than only posts from a single community.[5] The Home feed has important implications for this analysis, as I estimate the effect a post's rank in the subreddit feed has on its future engagement and how alternative ranking algorithms on the subreddit feed impact engagement patterns. Importantly, this captures both the direct effect of changing a post's rank on the subreddit feed and the indirect effect of changing the rank on the Home feed, holding fixed posts from other communities. For example, if two posts from the politics feed (A and B) and one post from another community (C) are ranked A, C, B on the Home feed, then counterfactuals where post B is promoted on the politics community correspond to the counterfactual ranking B, C, A on the Home feed. Given the prominence of the Home feed, it is important that the position effect estimates and counterfactual analyses include the effect of post rank in the Home feed.

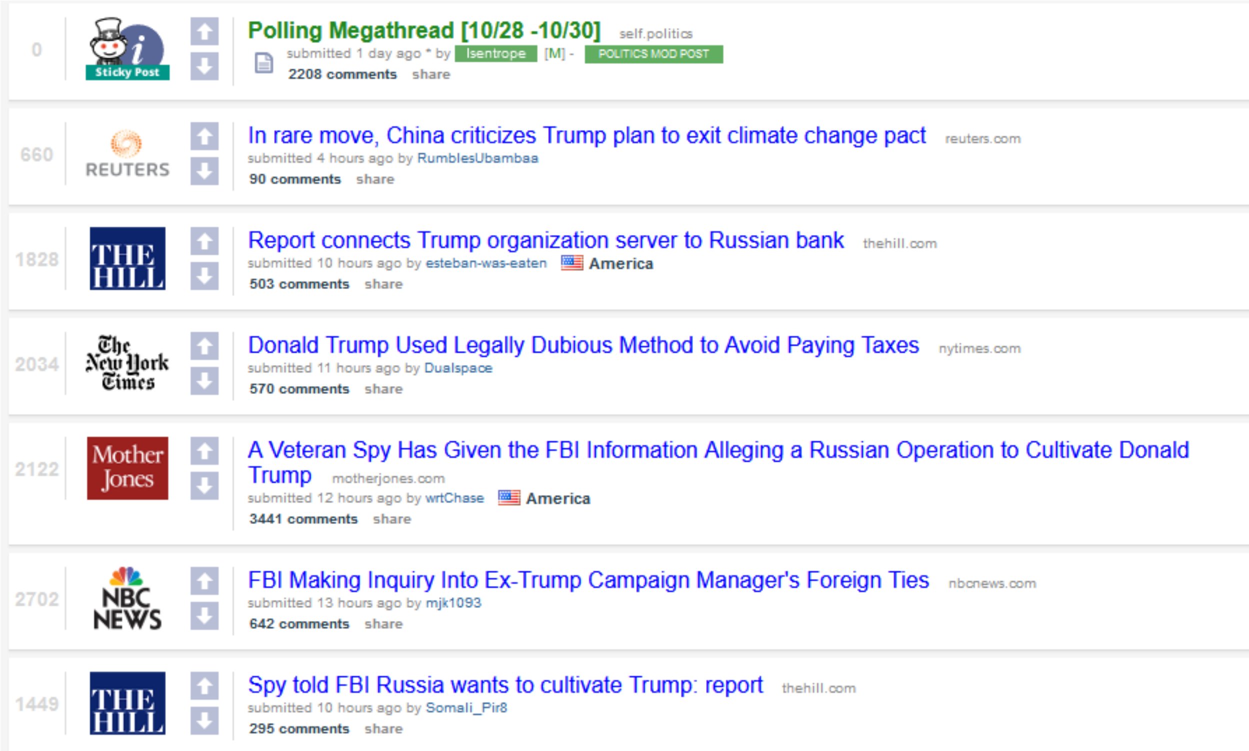
1.2.2 Data

Ranking and Engagement

I merge data from several sources in this study. First, I scrape subreddit landing pages from the Internet Archive's Wayback Machine for each subreddit in the study. These data provides historical snapshots of subreddit feeds, allowing me to collect the top 25 ranked posts, their position in the feed, and post features. A snapshot from the politics community following the 2016 election is shown in Figure 1.1. Alongside the post position in the feed, parsing the Wayback Machine snapshots provides the age of a post, number of existing comments, vote score of each post (net number of upvotes minus downvotes), post title, and domain

[5] In 2018, after the period studied, Reddit changed the default algorithm used by the Home feed to the Best algorithm, as described in https://www.reddit.com/r/changelog/comments/7spgg0/best_is_the_new_hotness/.

Figure 1.1: Snapshot of Politics Community



0 Polling Megathread [10/28 -10/30] self.politics
submitted 1 day ago * by Isentrope [M] - POLITICS MOD POST
2208 comments share

660 In rare move, China criticizes Trump plan to exit climate change pact reuters.com
submitted 4 hours ago by RumblesUbambaa
90 comments share

1828 Report connects Trump organization server to Russian bank thehill.com
submitted 10 hours ago by esteban-was-eaten America
503 comments share

2034 Donald Trump Used Legally Dubious Method to Avoid Paying Taxes nytimes.com
submitted 11 hours ago by Dualspace
570 comments share

2122 A Veteran Spy Has Given the FBI Information Alleging a Russian Operation to Cultivate Donald Trump motherjones.com
submitted 12 hours ago by wrtChase America
3441 comments share

2702 FBI Making Inquiry Into Ex-Trump Campaign Manager's Foreign Ties nbcnews.com
submitted 13 hours ago by mjk1093
642 comments share

1449 Spy told FBI Russia wants to cultivate Trump: report thehill.com
submitted 10 hours ago by Somali_Pir8
295 comments share

Note: A Wayback Machine snapshot of the politics subreddit from November 2016. Posts pinned by moderators are shown in green and are typically threads created to discuss major events or frequent discussion topics such as polling. Posts in blue are algorithmically ranked organic posts that are the focus of this study.

the post links to, if any. In addition, each snapshot reveals the number of subscribers each community has and the number of users online at the time of the snapshot.

Submissions on Reddit are either pinned to the top of the feed by community moderators or ranked organically.[6] I will focus on organic posts displayed in blue in Figure 1.1. These posts are submitted by users and ranked according to the algorithms described in Section 1.2.1. In the politics community considered here, posts are required to follow strict community guidelines: they must be on topic for the community, they must link to an article from a news publisher, and the post title must exactly match the headline of the article to which the post links. Any commentary on the article must be added in the comment sections, which I turn to next.

The primary engagement metric I consider is comments on articles. Reddit is a platform centered around sharing and discussing user-generated content. Comments themselves are a

[6] Posts pinned by moderators are shown in green and are typically threads created to discuss the major events of the day. Importantly, these are not algorithmically ranked and I condition on these posts remaining in their position on the feed. That is, I only consider counterfactuals where the organic post positions change.

form of user-generated content that bring people to the platform, and encouraging additional comments is of direct interest to the platform [Burke et al., 2009]. In addition, experimental evidence suggests users who receive comments on their posts are more likely to generate content in the future, a finding that further supports the premise that encouraging more comments is desirable for Reddit [Eckles et al., 2016, Mummalaneni et al., 2022]. Moreover, there is correlational evidence that users who comment more also spend more time on social media platforms – a metric that more closely approximates the amount of advertisements the user sees [Wojcik and Hughes, 2019]. I merge data from Baumgartner et al. [2020] that contain a near-universe of submissions and comments to public Reddit communities to generate the engagement outcomes. These data contain user-level commenting behavior, where each comment includes a time-stamp, a user identifier of the comment author, the full text of the comment, the post the comment is responding to, and the vote-score the comment received, among other observables. This information allows me to reconstruct a post's full comment history, including the comments that occurred immediately following each of the Wayback Machine's snapshots.

These data on user comments serve several purposes. First, they allow me to construct the number of comments each post received in a window following each snapshot. This will be critical for estimating position effects on the platform. Second, individual-level comment decisions are used to estimate a choice model of user engagement in Section 1.4. Here, the panel nature of the data allows me to identify rich user-level heterogeneity in comment preferences. Finally, studying comments allows me to analyze the text content to provide additional insight into user preferences and to understand how optimizing for engagement impacts the sentiment of comments submitted to the platform.

Publisher Ratings

I also collect publisher ratings that capture various aspects of an article's publisher. I use two sets of ratings. First are measures of a publisher's political slant [Robertson et al., 2018] that represent the relative propensity of a publisher domain being shared on Twitter by known Democratic party members relative to known Republican members, ranging from -1 to 1. A slant rating of -1 represents a domain that is only shared by Democrats while a slant rating of 1 represents a domain that is only shared by Republicans. Robertson et al. [2018] demonstrate this measure is consistent with a number of other expert, crowd-sourced, and audience-based ratings [Bakshy et al., 2015a, Budak et al., 2016]. A primary benefit of the Robertson et al. [2018] scores compared to other measures of publisher slant is the high coverage, as the data set includes ratings for over 19,000 domains. This results in high coverage in our data, with over 90% of posts in the politics community containing a link to a publisher matching a domain in the Robertson et al. [2018] data. The coverage is lower for other news categories, as some categories have less strict rules around sharing news articles and allow links to smaller websites such as sports blogs. The politics community, however, is the focus of this study and the other categories are only used to improve power in identifying position effects.

I also use credibility ratings, described in Lin et al. [2022], for over 11,520 news publishers. Lin et al. [2022] aggregate individual ratings from six rating organizations and demonstrate substantial agreement among individual sources. Importantly, the ratings released alongside Lin et al. [2022] show an extremely high correlation with NewsGuard ratings, a proprietary set of publisher ratings that employ extensive criteria including accuracy and balance of reporting, a process of publishing corrections, clear separation of opinion articles, and transparency of perceived conflicts. Figure 1.2 plots the joint distribution of publisher slant score and credibility rating for publishers that appear in at least 1% of the snapshots in the politics community. Table A.1 shows these ratings for six example domains. In evaluating user news diets, I will discretize the credibility ratings into high- and low-quality publishers for ease of interpretation. When doing so, I classify publishers as high quality if their credibility rating is greater than 0.65 and I show robustness of key results to other thresholds in Appendix Section A.3.4.[7]

1.2.3 Textual Analysis of Comment Text

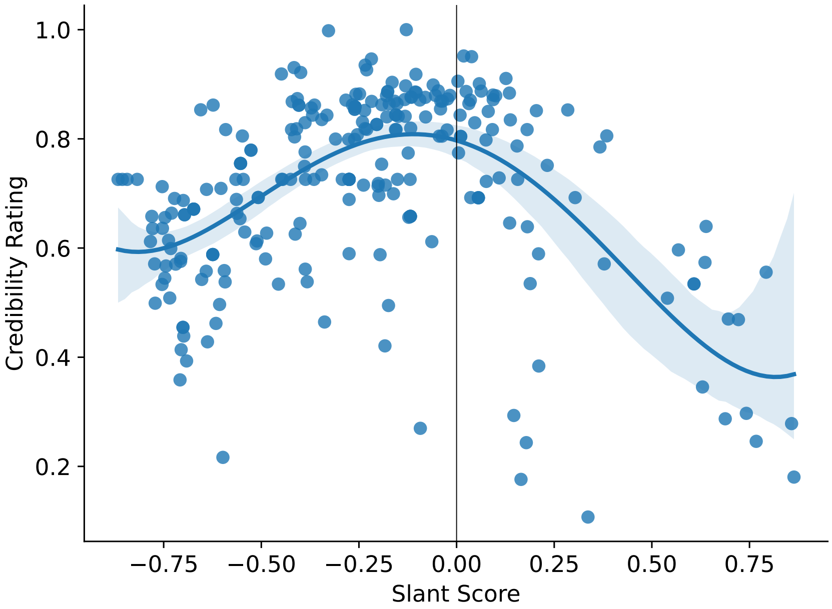
A unique benefit of studying comments as the focal measure of engagement is that I can analyze the textual content in order to understand the types of comments users are submitting to the platform and how this varies depending on article features. This sentiment analysis is used in the model of user engagement decisions as I allow users to choose the sentiment of their comment conditional on article features. I use these estimates to evaluate the extent to which optimizing for engagement leads to deterioration in discussion quality.

I analyze the sentiment and emotional content of comments using a pre-trained neural network for sentiment analysis and emotion detection. The pre-trained neural network, which is described in Pérez et al. [2021], uses embeddings generated using the language model BERTweet [Nguyen et al., 2020]. For each comment in the data, this model constructs a set of scores for predicted sentiment and emotion.

I focus on comment sentiment as the primary quality measure of a comment's text. Appendix Figure A.1 shows the correlation between text sentiment, predicted toxicity, and the comment's emotional content. Negative-sentiment comments are more likely to be classified

[7] The threshold of 0.65 is chosen as it is the median Lin et al. [2022] credibility rating within the Medium credibility category of Media Bias Fact Check, a professional rating organization.

Figure 1.2: Joint Distribution of Publisher Credibility and Slant Scores



Note: This figure plots the joint distribution of publisher slant score and credibility rating for the set of publishers that appear in at least 1% of the snapshots in the politics community. The dotted line displays the cutoff for high-credibility publishers. The regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Table 1.1: Summary Statistics

	Number	Share of Posts Missing			Number of Comments			
	Snapshots	Domain	Slant	Credibility	5 Min	10 Min	20 Min	60 Min
Politics	2105	0.01	0.07	0.11	2.98	5.97	11.93	35.76
US/World	5735	0.00	0.07	0.14	2.10	4.21	8.43	25.22
Sports	6390	0.22	0.66	0.85	0.78	1.53	3.01	8.51
Entertainment	3450	0.21	0.53	0.69	0.61	1.21	2.43	7.23
Gaming	2080	0.31	0.70	0.95	0.47	0.95	1.90	5.67
Technology	5912	0.06	0.43	0.66	0.28	0.57	1.13	3.41
Crypto	1267	0.38	0.77	0.87	0.23	0.45	0.90	2.60
Science	3561	0.10	0.37	0.49	0.10	0.20	0.41	1.20
Business	1632	0.08	0.27	0.38	0.05	0.09	0.19	0.58

Note: Summary statistics for the communities included in the study. Each row represents a category of news. The Number of Snapshots column contains the number of Wayback Machine snapshots for all communities in each category. The columns labeled Share of Posts Missing denote the share of submissions that lack information on publisher domain, slant score, and credibility rating. The columns labeled Number of Comments show the average number of comments a submission receives in the 5, 10, 20, and 60 minute periods following a snapshot. These columns average over both periods (i.e. snapshots) and positions in the feed.

as toxic, more likely to contain strong negative emotions such as disgust and anger, and less likely to exhibit joy. Moreover, inspection reveals comments labeled as negative by the Pérez et al. [2021] model are often extremely vulgar and unlikely to contribute productively to the discussion.

1.3 Estimating Position Effects

In this section, I estimate the causal effect post rank has on the number of comments the post receives. Recall I ultimately want to understand how engagement-maximizing ranking algorithms impact the type of content to which users are exposed and with which they engage. A key challenge in this analysis is the endogeneity of post rank, where the rank of a post is correlated with its potential outcomes. This section introduces the identification strategy I use to overcome this challenge by estimating position effects – the causal effect of post rank on engagement. This serves two purposes. First, the treatment effect estimates provide important motivation for the analysis, given that I find post rank has a large causal effect on engagement, meaning the ranking algorithm plays an important role in shaping the posts with which users eventually engage. Second, the causal estimates from this section will be utilized directly in identifying the choice model of user engagement that is employed in the analysis of counterfactual ranking algorithms.

Naive comparisons between posts with lower ranks (higher on the page) and posts with higher ranks (lower on the page) are unlikely to identify the causal effect of rank as I expect potential outcomes to be correlated with post rank [Narayanan and Kalyanam, 2015, Ursu, 2018]. It is likely that posts with high potential outcomes, or latent commentability, are more likely to be shown higher on the page. This dependence would be severed if Reddit ranked posts in random order and the effect of rank on engagement could be identified using the simple comparison [Ursu, 2018]. However, this is rarely the case in observational settings containing ranked content like the one studied here.

Therefore, I exploit a regression discontinuity to identify the causal effect of rank on engagement. Until 2017, Reddit maintained an open-sourced mirror of its code base, which allows me to directly inspect the algorithm used to sort posts [reddit.com, 2017]. The algorithm assigns a ranking score to each post and ranks posts in descending order of these scores. Formally, a post's ranking score is defined as

$$s_{jt} = \text{sign}\left(u_{jt}\right) \log_{10}\left(\max\left\{\left|u_{jt}\right|, 1\right\}\right) - \frac{a_{jt}}{45,000} \tag{1.1}$$

where u_{jt} is the net number of upvotes minus downvotes that post j had at time t and a_{jt} is the age of the post in seconds.[8] This requires that for the ranking score of a post with a positive vote score to remain constant, every 12.5 hours the net number of upvotes minus downvotes must increase by a factor of 10 to offset the age penalty. Importantly, this defines a continuous score that determines post rank, creating a regression discontinuity that can be used to identify position effects [Narayanan and Kalyanam, 2015].

To give a concrete example of the regression discontinuity I exploit, consider two adjacent posts i, j with ranking scores s_i, s_j and observed ranks r_i, r_j. There is a discontinuous jump in the probability of post i being ranked lower than post j when the continuous forcing variable $s_i - s_j$ crosses zero. I take advantage of this discontinuity to identify the effect of rank on future engagement, under the assumption that potential outcomes (i.e. latent post quality) are continuous across the zero threshold of the forcing variable $(s_i - s_j)$.

1.3.1 Implementation Details of the Regression Discontinuity Design

I now discuss the implementation details of the regression discontinuity used to estimate the causal effect of post rank on future engagement. In particular, I focus on estimating the causal effect of moving up from position $r+1$ to position r on the feed. To simplify notation,

[8] I normalize the post's timestamp by the period to interpret the second term as age. This is equivalent to adding a constant to all posts in a period and does not affect the ranking, but does make the ranking score more interpretable.

let D_i be a treatment indicator (i.e. $D_i = 1[s_i > s_{-i}]$), and the forcing variable is denoted $\Delta s_i = s_i - s_{-i}$.

In this setting, the running variable is a composition of two scores, the ages and vote scores of the posts. This creates a cutoff frontier, shown in Appendix Figure A.6, analogous to geographic regression discontinuity designs. I take advantage of the multiple score nature of the problem and estimate the treatment effect at the origin, which ensures that posts are balanced on both post age and vote score, as described in Cattaneo et al. [2023].

The primary results use a local-linear approximation to the conditional expectation functions on either side of the discontinuity and a uniform kernel. I will show the estimates are similar under alternative specifications. I restrict to observations within a bandwidth λ of the cutoff chosen to minimize the mean squared error of the treatment effect estimator [Calonico et al., 2014, Cattaneo et al., 2020] and demonstrate the results are not sensitive to this choice (Appendix A.2.2).

For each of the 24 positions on the first page of the feed, I estimate the treatment effect of moving from position $r+1$ to position r as

$$\hat{\tau}_r = \hat{\mu}_r^+ - \hat{\mu}_r^- \tag{1.2}$$

where $\hat{\mu}_r^+$ is the estimated intercept from the local linear regression to the right of the discontinuity and $\hat{\mu}_r^-$ is the intercept to the left of the discontinuity [Cattaneo et al., 2020]. I estimate the treatment effect separately for each position in the feed, allowing the treatment effect of being promoted from position $r+1$ to position r to vary by position. In Table 1.2 and Appendix A.2.2, I show the results are not sensitive to the degree of polynomial approximation or the choice of kernel. Following best practices [Cattaneo et al., 2020], statistical inference uses robust bias-corrected standard errors that are clustered at the period level.

Measurement Error in the Running Variable

A challenge in this setting is that the running variable is constructed using data scraped from the Internet Archive's WayBack Machine and the reconstructed ranking scores do not completely determine the rank of a post. This is a result of several factors. First, Reddit explicitly adds noise to the vote scores shown to users to combat vote manipulation [Muchnik et al., 2013].[9] Second, Reddit caches votes and rankings for performance purposes due to the large amount of traffic the platform receives.[10] Caching means the ranking score and actual

[9] https://www.reddit.com/wiki/faq

[10] https://web.archive.org/web/20170121192832/https://redditblog.com/2017/1/17/caching-at-reddit/

ranks are not continuously updated. This makes it possible for the observed ranks to differ from what is implied by the relative ranking scores as either the scores or observed rankings are a cached version.

Adding noise to the vote score introduces measurement error into the running variable and can bias traditional regression discontinuity estimates. To estimate the local average treatment effect of rank in the presence of measurement error in the running variable I follow Dong and Kolesár [2021] by excluding posts within a doughnut around the discontinuity. I manually select a doughnut width of 0.05 on either side of the cutoff and show in Appendix A.2.2 that the results are robust to the choice of doughnut width. Under the assumption that the doughnut excludes all periods where the posts are misclassified due to measurement error, Dong and Kolesár [2021] show that the usual regression discontinuity estimators identify a local average treatment effect.

After excluding posts within a doughnut of the discontinuity, I assume the remaining mismatch between post rank and the relative ranking scores is not due to measurement error. This assumption appears justified, as the probability of mismatch is constant as one moves away from the discontinuity (Figure A.2). If this were driven by the noise added to vote scores, the probability of mismatch would decline further away from the discontinuity as the probability the noise added is sufficiently large to misclassify the posts declines. Therefore, estimating the local average treatment effects using local linear regression results in conservative estimates of position effects.

1.3.2 Testing the Validity of the Regression Discontinuity Design

I now show evidence that Reddit ranks posts according to the algorithm I describe to establish a first stage in the regression discontinuity analysis. For each position on the feed $r \in \{1, \ldots, 24\}$, I consider the two posts ranked in position r and $r + 1$ and plot the probability that a post is in position r against the running variable (the difference in ranking scores of the two competing posts). Figure 1.3a shows the discontinuity between position 1 and position 2; the plots for the remaining positions are shown in Appendix A.2.1. There is a clear discontinuity in the probability of being ranked lower when a post's ranking score surpasses the competing post's ranking score in a period.

In addition, to test that post observables are balanced across the discontinuity, Figure A.11 plots the estimated treatment effect of rank on pre-treatment covariates including publisher slant, publisher credibility, post vote score, and post age. Nearly all estimates are insignificant at the 5% level, suggesting post observables are balanced across the discontinuity. While it is not possible to test the identifying assumption that potential outcomes are continuous through the discontinuity, this result is consistent with such an assumption holding. Appendix A.2.2 presents plots of the non-parametric conditional expectation functions of these covariates around the discontinuity.

1.3.3 Position Effect Estimates

I now turn to estimating how post rank affects the engagement a post receives in the window following the snapshot. Figure 1.3b plots a binned scatter plot of the log of one plus the number of comments a post receives in the 20 minutes following the snapshot against the running variable (Δs_i) to visualize the discontinuity in the outcome variable. There is a clear discontinuity in engagement when a post is promoted to position 1 from position 2. Appendix A.2.1 shows the same plots for the remaining positions on the feed and the discontinuity in engagement quickly disappears further down the feed, suggesting treatment effects of rank are largest at the top of the feed.

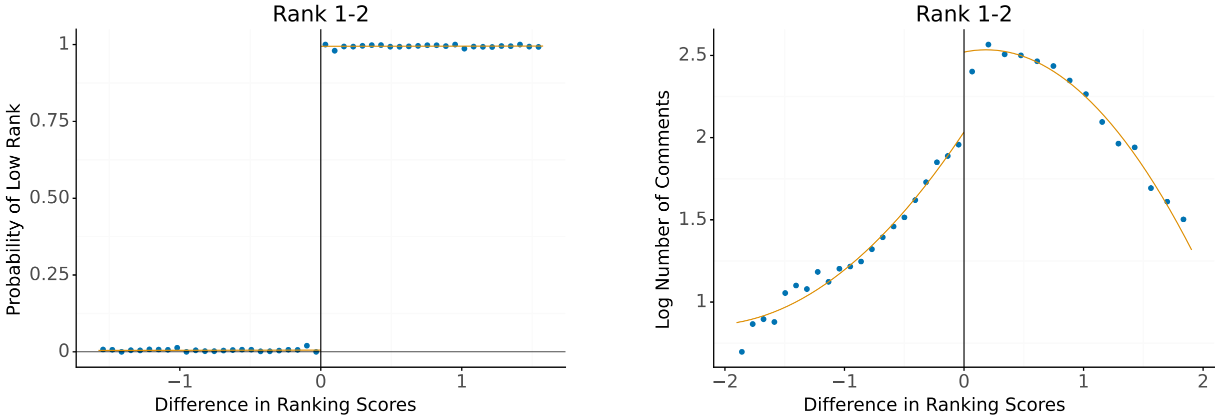
I estimate the local average treatment effect using local linear regression, and present treatment effect estimates in Table 1.2, which shows the effect of moving from rank $r + 1$ to rank r on the log of one plus the number of comments a post receives in the 20 minutes following a snapshot. Being promoted to the first position has a large effect, with the treatment effect estimate suggesting a 42.5% increase in the number of comments received immediately following a snapshot relative to the second post. The importance of rank quickly dissipates further down the feed. Table 1.2 also shows the results are robust to the polynomial choice and choice of kernel. Table 1.2 includes naive OLS estimates of position effects. As expected, OLS substantially overestimates the effect of position on engagement, and this is particularly severe towards the top of the feed.

These treatment effect estimates demonstrate that the ranking algorithm has an important effect in determining the posts with which users engage. This in turn motivates further investigation of what content is promoted when rankings are designed to optimize for engagement, since the platform has substantial power in determining what content users are exposed to and ultimately engage with.

Figure 1.3: Regression Discontinuity Plots

(a) Effect on Post Rank

(b) Effect on Engagement



Note: Regression discontinuity plots for the discontinuity around being promoted to the top position on the feed from the second position on the feed. Here, the x-axis is the running variable – the difference in the focal post's ranking score from that of the adjacent post – and the y-axis is (a) the probability a post is ranked lower on the page and (b) the log number of comments received in the 20 minutes following a snapshot plus one. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial fits are plotted alongside the binned mean values. The corresponding figures for the remaining positions on the feed are shown in Appendix A.2.1.

Table 1.2: Position Effect Estimates

		Regression Discontinuity		
Rank	OLS	Local Linear	Local Constant	Triangular Kernel
1	0.904	0.354	0.208	0.365
	(0.012)	(0.142)	(0.318)	(0.164)
2	0.249	0.244	0.204	0.250
	(0.010)	(0.045)	(0.093)	(0.050)
3	0.145	0.156	0.120	0.160
	(0.009)	(0.033)	(0.068)	(0.037)
4	0.097	0.123	0.128	0.113
	(0.009)	(0.032)	(0.066)	(0.036)
5	0.073	0.107	0.088	0.111
	(0.009)	(0.038)	(0.079)	(0.042)
6	0.067	0.116	0.140	0.106
	(0.008)	(0.037)	(0.077)	(0.040)
7	0.056	0.077	0.062	0.067
	(0.008)	(0.035)	(0.074)	(0.039)
8	0.043	0.071	0.077	0.052
	(0.008)	(0.036)	(0.075)	(0.040)
9	0.043	0.027	0.004	0.028
	(0.007)	(0.031)	(0.064)	(0.034)
10	0.039	0.063	0.075	0.065
	(0.007)	(0.025)	(0.051)	(0.027)
11	0.040	0.055	0.079	0.054
	(0.007)	(0.027)	(0.056)	(0.030)
12	0.016	0.047	0.054	0.054
	(0.007)	(0.035)	(0.074)	(0.038)

Note: Estimates of the local average treatment effect from a post moving from position $r+1$ to position r on the feed on the log of one plus the number of comments a post receives in the 20 minutes following a snapshot. Robust bias-corrected standard errors that allow for misspecification of the conditional expectation function and that are clustered at the period level are shown in parentheses. Estimates exclude posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The bandwidths for each rank are not varied across the various regression discontinuity specifications to isolate the difference due to the different specifications. The corresponding table for positions 13-24 is shown in Appendix A.2.1.

1.4 Model of Individual Engagement Decisions

I now estimate a model of user engagement that allows me to estimate engagement patterns under counterfactual ranking algorithms. Section 1.4.1 introduces the model of individual decisions, Section 1.4.2 describes identification and the estimation approach, and Section 1.4.3 summarizes the model estimates and fit.

1.4.1 Model

Users indexed by i visit the platform in periods indexed by t and are exposed to a ranked feed of posts indexed by j. In each period, users are exposed to a post in position r_{jt} if $v_{ijt} = 1$, which is an independent Bernoulli random variable equal to one with probability $p\left(r,t\right)$. I use a parsimonious parameterization of the exposure probability, $p\left(r,t\right) = p_t p_r$, where p_t is the probability of accessing the platform in period t and p_r is the probability of being exposed to a post in position r conditional on accessing the platform. Section 1.5.4 endogenizes the search process and demonstrates the main findings are not sensitive to allowing users to reallocate attention across the feed in equilibrium. If exposed to a post, users receive utility

$$u_{ijt} = \delta_{ijt} + \varepsilon_{ijt} \tag{1.3}$$

if they comment on post j in period t, which I denote d_{ijt}. Consumers comment if they are exposed to the post and the utility from commenting exceeds the utility from the outside option (u_{i0t})[11]

$$d_{ijt} = 1\left[v_{ijt} = 1\right] 1\left[u_{ijt} \geq u_{i0t}\right]. \tag{1.4}$$

I model $\delta_{ijt} = x'_{jt}\left(\bar{\beta} + \beta_i\right) + \xi_{jt} = \delta_{jt} + x'_{jt}\beta_i$ where x_{jt} is a vector of observable article features, $\bar{\beta}$ represents average preferences, and β_i is a vector of the deviation of user i's preferences from the mean. Finally, ξ_{jt} is a article-period fixed effect that represents latent article commentability. This fixed effect captures anything users have vertical preferences over. For example, if users prefer to comment on articles with many existing comments, this

[11] A key simplification is the stylized process by which users form consideration sets. A more flexible model that allows users to consider a subset of posts and comment on their most preferred post introduces substantial computational challenges as the number of potential consideration sets grows combinatorially. These models would likely yield similar results given users are highly unlikely to comment on more than one post in either the data or in the counterfactual simulations. Therefore, given the substantial computational benefits of assuming independence between comment decisions, I assume that users make engagement decisions on posts independently. The independence assumption between posts also prohibits preferences that depend on features of other posts on the feed such as a preference for a diverse feed. Experimental evidence from a music recommender system found little evidence of a preference for diversity consistent with the approach taken here [Chen et al., 2023].

would load onto the ξ_{jt} term.[12] Post rank is excluded from utility in this model, implying that rank does not impact choice conditional on exposure to a post.[13] Finally, normalize $\delta_{i0t} = 0$ and assume ε_{ijt} is an independent and identically distributed Type 1 Extreme Value preference shock. This results in the mixed logit choice probabilities multiplied by the exposure parameter $p(\cdot)$.

$$P_{ijt} = P\left(v_{ijt} = 1, u_{ijt} \geq u_{i0t}\right) = p\left(r_{jt}, t\right) \frac{\exp \delta_{ijt}}{1 + \exp \delta_{ijt}} \tag{1.5}$$

Conditional on commenting on a post, users choose the sentiment of the comment to submit. Users can either submit a comment with negative sentiment or neutral sentiment. Users choose the probability with which their comment will be perceived negatively based on the user-specific and vertical components of comment utility. That is, conditional on commenting users choose the probability that comment ijt will be a negative comment as follows

$$\log \frac{b_{ijt}}{1 - b_{ijt}} = \beta^s_{i0} + \beta^s_{i1}\left(\delta_{ijt} - \xi_{jt}\right) + \beta^s_{i2}\xi_{jt} + \varepsilon^s_{ijt} \tag{1.6}$$

where b_{ijt} is the probability user i's comment on post jt is a negative comment, $\delta_{ijt} - \xi_{jt}$ is the user-specific component of comment utility user i receives when commenting on post jt, ξ_{jt} is the vertical commentability component of post jt, $\beta^s_i = \langle \beta^s_{i0}, \beta^s_{i1}, \beta^s_{i2} \rangle$ is a vector of individual i's sentiment preferences, and ε^s_{ijt} is an independent error term.

1.4.2 Identification and Estimation

Identification of Model Parameters

The key identification challenge in this model is that observed post ranks are correlated with latent commentability, or $E\left[\xi_{jt} r_{jt}\right] \neq 0$. I now describe how this model is identified using the regression discontinuity from Section 1.3 and user-level engagement decisions.

I first describe how the exposure parameters (p_t, p_r) are identified. I assume that each user logs on to the platform with probability p_t, independent of article features or preference shocks. This probability is identified via the share of users who visit the platform in each

[12] The latent commentability term ξ_{jt} is often referred to as latent quality in the literature estimating demand systems. To avoid confusion with publisher credibility, I refer to ξ_{jt} as latent commentability, where this captures a vertical component making all users more likely to comment on the article.

[13] This assumption is motivated by the findings of Ursu [2018] that demonstrates empirically that rankings impact search probabilities but, conditional on search, do not affect purchase probabilities in an online travel platform. This is also consistent with recent work modeling personalized rankings in e-commerce [Donnelly et al., 2023]

period which I estimate using data on the share of users online at the start of each period. Conditional on accessing the platform, I assume all users are exposed to the top post on the feed implying that $p_1 = 1$.[14] The remaining exposure parameters p_r are identified by the reduced form treatment effects after I impose constant treatment effects[15]

$$\tau_r = \log \frac{E\left[d_{ijt}(r)\right]}{E\left[d_{ijt}(r+1)\right]} = \log \frac{p_r}{p_{r+1}}. \tag{1.7}$$

The assumption of constant treatment effects could in principle be relaxed to allow for arbitrary individual heterogeneity and heterogeneity along observed article features, though the demands on the data grow substantially if this type of heterogeneity are included. For example, allowing for arbitrary individual heterogeneity would require estimating the reduced form treatment effects separately for each user.

Given exposure parameters, individual preference parameters and mean preference parameters are identified from the assumption that article features are exogenous $E\left[x_{jt}\varepsilon_{ijt}\right] = 0$ and $E\left[x_{jt}\xi_{jt}\right] = 0$. Finally, the parameters in the sentiment model are identified by the assumption that ε^s_{ijt} is mean-independent of ε_{ijt}, ξ_{jt}, and x_{jt}.

Estimation

I estimate the choice model using data from the politics community given the relevance of this community to managers, policy makers, and users. I use individual-level comment decisions and restrict the sample to users who comment on at least 25 articles in the periods I study, where a period consists of the 60 minutes following a Wayback Machine snapshot. In Appendix A.4, I show the results are similar using an approach that trains a recommender system on a larger and more representative sample of users, thereby providing confidence that the results generalize to a broader set of users. Nonetheless, this sample of highly active users is also of direct interest to the platform because user-generated content is vital to platform's business model.

I take this model to the data using a two step procedure that simplifies the computation given the large number of periods, users, and posts. In the first step, I estimate the exposure parameters p_t and p_r. I estimate p_t by combining data on the number of users online at the start of each period, which are observed in the Wayback Machine snapshots, with public statements by the platform on the average session duration to estimate the number of users who log on to the platform during each period using Little's law [Little, 1961]. I then use

[14] The results are robust to other choices of p_1 as shown in Appendix Section A.3.4.

[15] That is, I assume $E\left[Y_{jt}(1)\right] = e^{\tau_r} E\left[Y_{jt}(0)\right]$ where $Y_{jt}(D)$ is the potential outcome under treatment D for the number of comments post j in period t received following a snapshot.

public usage statistics again to estimate the number of active community members and calculate the share of active community members who log on in each period. Finally, I smooth estimates of p_t by taking the fitted values of a regression of the raw values of p_t on quarter and day of week fixed effects. The full details of this process are described in Appendix A.3.3. I then estimate the remaining exposure parameters using an empirical analog of Equation 1.7 ($p_r = \exp\left\{-\sum_{r'>1} \tau_{r'-1}\right\}$).

Second, given the estimates of p_t and p_r I estimate the individual preference parameters β_i using maximum likelihood

$$\mathcal{L} = \sum_t \sum_i \sum_j d_{ijt} \log P_{ijt} + (1 - d_{ijt}) \log (1 - P_{ijt}) . \tag{1.8}$$

Finding the maximum likelihood estimate involves solving a high-dimensional optimization procedure due to the large number of individuals and posts. Therefore, I use the following iterative algorithm. First, I initialize a guess of ξ_{jt} and, conditional on these unobserved commentability parameters, estimate the individual-level preference parameters using maximum likelihood. I then invert observed engagement shares [Berry, 1994] using the Berry et al. [1995] contraction mapping to find the values of ξ_{jt} such that predicted market shares equal observed market shares.[16] I iterate between these two steps until convergence. Splitting the estimation algorithm into these two steps allows the maximum likelihood parameters to be estimated in parallel vastly reducing the required computation. Inference on the preference parameters uses cluster-robust standard errors.

The preference estimates of any individual user will contain substantial sampling error given the limited number of periods and comment decisions. This implies that the distribution of preference estimates will be a convolution of the true distribution of preferences and sampling error, leading the distribution of estimates to be over-dispersed relative to the true distribution. To correct for this over dispersion, I shrink all preference estimates towards the grand-mean using the empirical Bayes procedure described in Appendix A.3.2.

1.4.3 Model Estimates

To assess the fit of the model, Table 1.3 presents summary statistics of actual engagement and engagement predicted by the model. The distribution of actual engagement with engagement predicted by the model is also shown graphically in Figure A.17. Figure A.17a demonstrates the correlation between actual user engagement and predicted user engagement. The correlation is high, though the model tends to overestimate total engagement

[16] There are instances in the data where a post receives zero comments. I assume that ξ_{jt} is bounded below such that the minimum predicted market share is equal to 0.01% in these situations.

for users with the highest engagement. Figure A.17b shows the correlation between actual and predicted user engagement by publisher credibility and the model again has a high correlation between actual and predicted engagement by group.

Table C.32 summarizes the distribution of individual preference estimates $(\bar{\beta} + \beta_i)$. It is helpful to summarize preferences for publisher slant through each user's bliss point, which is defined as the slant the user most prefers

$$b_i^* = \begin{cases} \text{sign}\left(\beta_{is}\right) & \text{if } \beta_{is^2} \geq 0 \\ \min\left\{1, \max\left\{-1, -\frac{\beta_{is}}{2\beta_{is^2}}\right\}\right\} & \text{if } \beta_{is^2} < 0 \end{cases} \tag{1.9}$$

where β_{is} (β_{is^2}) is user i's taste parameter on post slant (slant squared). The marginal distribution of slant bliss points is shown in Figure 1.4. It is evident there is substantial heterogeneity within preferences in the politics community. Just over half of users prefer more credible publishers, while the remaining users prefer less credible publishers. Regarding political slant, there is also substantial heterogeneity, with a large mass of users preferring outlets slightly left of center. There are also mass points at each political extreme, with nearly 20% of users preferring outlets that are strongly left leaning and over 25% of users preferring outlets that are strongly right leaning.

Table 1.4b summarizes the estimates of individual-level preferences to submit a negative comment based on the vertical- and individual-specific components of comment utility (Equation 1.6). There is substantial heterogeneity in user preferences to submit a negative comment with 51.6% of users more likely to comment negatively on posts in which they are more likely to comment. Figure A.18 reveals this is especially true for users more likely to comment on left-leaning posts (i.e. they have a negative bliss point) and users who prefer to comment on less credible publishers. Ranking algorithms that optimize solely for engagement will increase the share of negative posts for these users.

Table 1.3: Summary of Model Fit

		Actual		Model	
		Mean	Std	Mean	Std
Total		52.39	38.85	54.05	35.31
Credibility	High	41.55	30.73	42.72	27.88
	Low	10.84	9.62	11.33	8.26
Slant Partition	Strongly Left	13.38	11.86	14.04	10.42
	Left	7.99	6.75	8.14	5.31
	Middle	13.44	10.60	13.63	8.88
	Right	13.63	10.78	13.93	9.15
	Strongly Right	3.95	3.87	4.31	3.32

Note: Summary of the model fit. The Actual columns report the average and standard deviation of the total number of comments posted by each user and the number of comments by publisher rating. The Model columns report the model's predicted values for the same quantities under the existing ranking algorithm.

1.5 Counterfactual Ranking Algorithms

1.5.1 Background on Engagement Based Feeds

While it is beyond the scope of this article to provide a comprehensive review of the architectures and implementations of news feed algorithms deployed in practice, I will give a high-level description that abstracts away from many of the low-level platform-specific implementation details. See Thorburn et al. [2022] and Narayanan [2023] for more thorough reviews. At a high level, social media ranking algorithms can typically be decomposed into two primary components: candidate generation and ranking.

In candidate generation, the algorithm usually selects a set of posts that are eligible to be shown to a user. As described in Thorburn et al. [2022], this is often either a computationally efficient algorithm that filters posts based on a user's network – examples include the Facebook News Feed [Lada et al., 2021] and Twitter Timeline [Twitter, 2023] – or a bare bones implementation of the ranking algorithm, as in the YouTube homepage [Covington et al., 2016]. Candidate generation also generally includes content moderation filters that remove posts deemed ineligible to be promoted.

Table 1.4: Distribution of Individual Preference Estimates

	Mean	Std	1%	25%	50%	75%	99%
Constant	-4.55	0.75	-6.09	-5.06	-4.62	-4.11	-2.49
Slant Score	-0.20	0.64	-1.60	-0.65	-0.20	0.23	1.27
Slant Score^2	-0.30	0.82	-2.10	-0.88	-0.30	0.27	1.59
Credibility Rating	-0.02	0.86	-2.30	-0.53	0.03	0.56	1.86
ξ_{jt}	-0.00	0.93	-2.22	-0.61	0.02	0.61	2.12

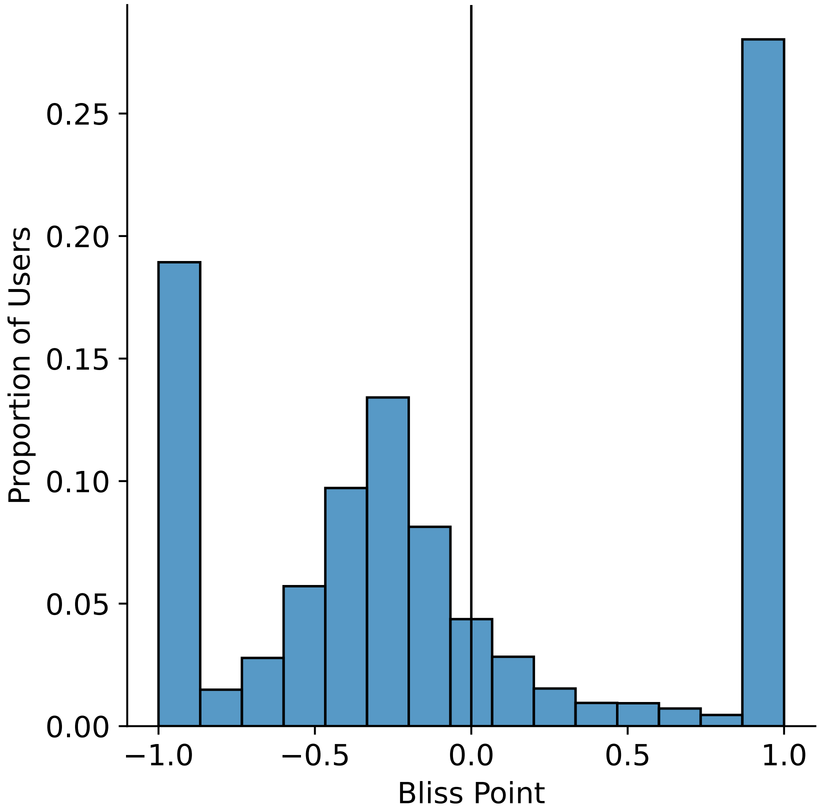
(a) Individual Comment Preference Estimates

	Mean	Std	1%	25%	50%	75%	99%
Intercept	0.20	7.25	-18.46	-3.88	0.20	4.26	18.78
Heterogeneous component	0.04	1.53	-3.93	-0.82	0.06	0.90	3.94
Vertical component	0.03	0.08	-0.19	-0.01	0.03	0.08	0.23

(b) Individual Sentiment Preference Estimates

Note: This table shows the user-level distribution of preference estimates. Panel (a) presents the distribution of comment preferences (Equation 1.4). The values for Constant, Slant Score, Slant Score^2, and Credibility Rating contain the user-level comment preferences. The values of ξ_{jt} are at the article level and show the distribution of latent article commentability. Panel (b) presents the distribution of sentiment preferences (Equation 1.6). The heterogeneous component captures how the likelihood of a user to submit a negative comment changes in response to changes in the user-specific component of post comment utility. The vertical component captures how the likelihood of a user to submit a negative comment changes in response to a change in the latent commentability term (ξ_{jt}). All preference parameters are shrunk to the grand mean using empirical Bayes.

Figure 1.4: Distribution of Slant Bliss Points



Note: This figure plots the marginal distribution of user-level slant bliss points. The bliss point is the slant score for which a user is most likely to comment, all else being equal. A bliss point of -1 implies the user is most likely to comment on left-leaning articles and a bliss point of 1 implies the user is most likely to comment on right-leaning articles.

In the ranking step, platforms typically employ a more complicated model that orders posts based on predicted engagement – often implemented as a weighted average of predicted clicks, time spent, comments, and votes [Lada et al., 2021, Thorburn et al., 2022, Twitter, 2023]. Additional higher-level signals such as predicted survey responses are occasionally included in the ranking objective function as well. Finally, the ranking step often includes a post-processing procedure that adjusts the ranking to avoid, for example, showing users only posts from a single highly engaging account.

1.5.2 Description of Counterfactual Rankings

In this study, I focus on the implications of different objective functions in the algorithmic ranking step conditional on candidate generation. Therefore, the counterfactuals considered here only re-rank the top 25 posts in each period, which I treat as the set of candidate posts to be ranked. This decision is relatively innocuous for analyzing the impact of optimizing for engagement, as latent post commentability is highly correlated with post rank in the data (Figure A.15). Latent post commentability is an important factor in optimizing for engagement, meaning posts that are not in the top 25 posts would be less likely to be ranked high on the feed even if they were included in the candidate posts.[17]

I assume that the platform has high quality estimates of user preferences given their access to rich user-level behavioral data and therefore assume the platform observes $\hat{\beta}_i$ in the counterfactuals. I assume the platform does not, however, observe latent article commentability (ξ_{jt}) and must estimate this through observable article features. I model the platform's estimates of ξ_{jt} as a supervised learning problem where the platform forms estimates of the true latent article commentability ($\hat{\xi}_{jt}$) based on article observables. I operationalize this using a random forest that predicts ξ_{jt} using observable post features including the stock of total and top-level comments, vote score, post age, publisher slant, and publisher credibility rating. This model performs well in the prediction task as demonstrated in Appendix Figure A.19, where it achieves an R^2 of 0.42. With estimates of article commentability, observed post features, and observed user preferences, the platform can estimate engagement probabilities for each user and article conditional on exposure $\hat{P}_{ijt} = \frac{\exp \hat{\delta}_{ijt}}{1+\exp \hat{\delta}_{ijt}}$ where $\hat{\delta}_{ijt} = x'_{jt}\left(\bar{\beta} + \beta_i\right) + \hat{\xi}_{jt}$.

With predicted engagement probabilities for each user-article, I estimate what content users would have engage with if the platform used the algorithms described below to re-rank posts according to observable post features and estimated engagement probabilities. To calculate engagement under a counterfactual algorithm, I calculate engagement probabilities for each post and user by multiplying the exposure probability for the post under the counterfactual ranking with the true estimated engagement probability conditional on exposure.

Non-personalized engagement maximizing: The non-personalized engagement maximizing algorithm solves the following maximization problem

$$r_t^N = \arg\max_{r\in\mathcal{R}} \sum_{j=1}^{J} p\left(r_j, t\right) E\left[\hat{P}_{ijt}\right] \tag{1.10}$$

where $\mathcal{R}$ is the set of possible rankings, $r = \langle r_1, \ldots, r_J\rangle \in \mathcal{R}$ is a vector of possible article ranks, and $\hat{P}_{ijt}$ is the platform's estimate of the probability that user i engages with article j in period t conditional on exposure. It is straightforward to show that when $p\left(r, t\right)$ is weakly

[17] This assumption does preclude analyzing simple proposed algorithms such as reverse chronological, as the restricted set of candidate posts excludes the high volume of low-quality posts that are often promoted under a reverse-chronological ranking.

decreasing in r, the optimal ranking sorts articles in descending order of $E\left[\hat{P}_{ijt}\right]$.[18]

Personalized engagement maximizing: The leading counterfactual considered is personalized engagement maximization. The personalized engagement-maximizing ranking solves

$$r_{it}^{P} = \arg\max_{r \in \mathcal{R}} \sum_{j=1}^{J} p\left(r_j, t\right) \hat{P}_{ijt} \tag{1.11}$$

which by a similar argument ranks articles in descending order of $\hat{P}_{ijt}$ for each user.

Credibility-aware algorithm: While short-term engagement is often used as a proxy for consumer welfare, a growing literature has emerged to study situations where these measures may differ. This disconnect can arise for rational economic agents [Spence and Owen, 1977] and in models with behavioral biases, including agents with present bias [Kleinberg et al., 2022], dual self models, [Agan et al., 2023, Kahneman, 2011], and digital addiction [Allcott et al., 2022]. Moreover, the platform may want to avoid promoting low-credibility publishers for brand-safety purposes or to prevent potential regulatory actions. These factors could lead the platform to consider publisher credibility in the ranking objective function. Therefore, I consider credibility-aware algorithm that maximizes an objective function that balances two competing objectives: total engagement and engagement with high credibility publishers

$$\mathcal{S}_{ijt} = E\left[d_{ijt}\right]\left((1-\lambda) + \lambda 1\left[c_{jt} \geq \underline{c}\right]\right) \tag{1.12}$$

where λ reflects the weight on engagement above a minimum credibility threshold $\underline{c}$. Note that this nests the personalized engagement-maximizing algorithm when $\lambda = 0$ and a credibility-maximizing algorithm when $\lambda = 1$. The credibility-aware algorithm solves

$$r_{it}^{O} = \arg\max_{r \in \mathcal{R}} \sum_{j=1}^{J} \hat{\mathcal{S}}_{ijt} = \arg\max_{r \in \mathcal{R}} \sum_{j=1}^{J} p\left(r_j, t\right)\left((1-\lambda) + \lambda 1\left[c_{jt} \geq \underline{c}\right]\right) \hat{P}_{ijt} \tag{1.13}$$

and is solved by ranking articles in descending order of $\left((1-\lambda) + \lambda 1\left[c_{jt} \geq \underline{c}\right]\right) \hat{P}_{ijt}$ for each user.

Benchmarks: I compare the engagement patterns under the counterfactual algorithms described above to two benchmark algorithms, the ranking employed by the platform (Ac-

[18] To show this, assume for contradiction there exists an optimal ranking with two posts j and j' such that $r_j < r_{j'}$ and $E\left[\hat{P}_{ijt}\right] < E\left[\hat{P}_{ij't}\right]$. Note that the objective under this ranking is weakly less than the objective if the positions of the two posts are swapped

$$\left(E\left[\hat{P}_{ij't}\right] - E\left[\hat{P}_{ijt}\right]\right)\left(p\left(r_j, t\right) - p\left(r_{j'}, t\right)\right) \geq 0$$

because $p\left(\cdot\right)$ is weakly decreasing in r. Therefore, this ranking is not optimal, thus providing a contradiction.

tual) and a random benchmark that randomly shuffles the articles shown on the page for each user (Random).

1.5.3 Counterfactual Ranking Algorithm Results

Summaries of engagement patterns under the different counterfactual ranking algorithms are shown in Table 1.5. I now turn to describing the quantity, credibility, diveristy, and sentiment of engagement in addition to studying how the various algorithms impact publisher market shares.

Impact on Engagement Quantity

The counterfactual analysis suggests that the algorithm employed by the platform, which prioritizes simplicity and transparency, is far from engagement maximizing. That said, the actual algorithm does substantially increase engagement relative to the random benchmark. As expected, optimizing explicitly for engagement leads to a substantial increase in engagement quantity. Much of the benefit comes from ranking articles according to expected engagement without personalization, which is evidenced by the 19.1% increase in engagement under the non-personalized engagement-maximizing algorithm. Personalizing user feeds increases engagement by 21.0% relative to the existing algorithm, providing a modest increase in engagement relative to the non-personalized engagement-maximizing algorithm. While modest in size, this lift does demonstrate the platform has an incentive to personalize rankings to drive engagement. Optimizing for engagement with high-credibility publishers also leads to a substantial increase in engagement relative to the actual algorithm employed (15.0%), but represents a substantial cost in terms of lost engagement relative to the engagement-maximizing algorithms.

Table 1.5: Counterfactual Engagement Summaries

	Engagement	Distance from Uniform	Max Partition Share	Credibility	Negative Engagement Share
Intercept	54.048	0.280	0.290	0.790	0.512
	(0.387)	(0.001)	(0.000)	(0.001)	(0.002)
Random	-6.468	-0.004	-0.001	-0.001	-0.001
	(0.043)	(0.000)	(0.000)	(0.000)	(0.000)
Non-Personalized	10.309	-0.008	-0.004	0.008	0.001
	(0.062)	(0.000)	(0.000)	(0.000)	(0.000)
Personalized	11.357	0.030	0.021	-0.001	0.002
	(0.072)	(0.001)	(0.000)	(0.000)	(0.000)
Credibility Maximizing	8.118	0.008	0.018	0.107	0.001
	(0.054)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	41675	41675	41675	41675	41675
R-Squared	0.031	0.031	0.084	0.405	0.000

Note: This table reports estimates of a panel regression of each counterfactual outcome on counterfactual algorithm dummy variables. The intercept is the average quantity under the existing algorithm. (1) Engagement represents the total number of articles a user comments on. (2) Diversity represents the first Wasserstein distance of engagement shares across publisher slant partitions from the uniform distribution. Recall distributions closer to uniform will have smaller distances, meaning they represent more diverse engagement. (3) Max partition share represents the max share of engagement in across publisher slant partitions. (4) Credibility represents the share of a users engagement with high-quality publishers. (5) Negative engagement share represents the share of comments that are negative sentiment. Standard errors are clustered at the user level.

Impact on Engagement with Publishers by Credibility

Reddit's algorithm does not materially impact the share of engagement with high-credibility publishers relative to a random ordering of posts, with both algorithms resulting in 79.0% of the average user's engagement being with high-credibility publishers. Optimizing for engagement also does not lead to a substantial change for the average user, with an average high-credibility engagement share of 79.9% for the non-personalized engagement-maximizing algorithm and 79.0% for the personalized engagement-maximizing algorithm. The credibility-maximizing algorithm does lead to a substantial increase in the share of engagement with high-quality publishers, with the average user's high quality engagement share rising to 89.7%.

Focusing on average changes masks important heterogeneity. Figure 1.5a plots the empirical CDF of the change in high-credibility shares relative to the existing algorithm. The non-personalized algorithm has little impact on the quality of news diets as the share of engagement with high-quality publishers does not change substantially for any user. The personalized engagement-maximizing algorithm, however, does have substantial impacts for many users despite the negligible average effect. The majority of users experience a modest improvement in the quality of their news diets, as a slightly larger share of their engagement is with high-credibility publishers. However, 41.5% of users experience a deterioration in the quality of their news diets, with a subset of these users seeing the share of their engagement with high quality publishers falling by over 10 percentage points. To better understand what users experience these declines, Figure 1.5b plots the relationship between news diet quality under the existing algorithm against news diet quality under the counterfactual algorithms. It is clear that users engaging with less credible publishers under Reddit's actual algorithm experience large declines in the quality of their news diets under the personalized engagement-maximizing algorithm. This suggests the engagement maximizing algorithm exacerbates differences in the quality of user news diets by promoting high-credibility publishers to the majority of users who typically engage with high-credibility publishers and promoting low-credibility publishers to users who have engaged with these publishers in the past. Moreover, the results are robust to the choice of threshold for high-credibility publishers (Appendix Section A.3.4) and I find personalization exacerbates differences in the quality of user news diets even for very low thresholds for high-credibility publishers.

Turning to the credibility-maximizing algorithm, I find that optimizing for engagement with high-credibility publishers leads to substantial increases in the share of engagement with high-credibility publishers across all users. Importantly, though, Figure 1.5b shows that the users experiencing the largest increases are those who engage more with low-credibility publishers under Reddit's actual algorithm. This indicates that including publisher credibility in the objective function narrows the disparity between users with high- and low-quality news diets, a difference that was exacerbated when optimizing only for engagement.

Impact on Engagement Diversity

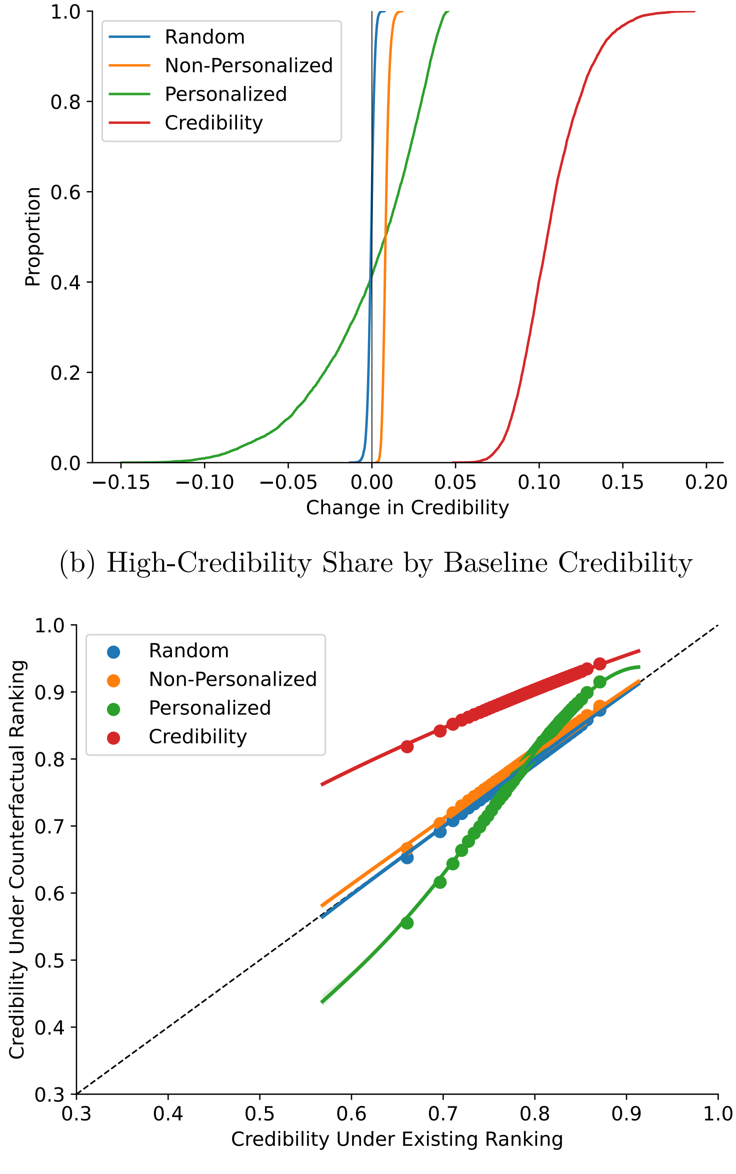
I now study the impact the counterfactual algorithms have on the diversity of engagement across the political spectrum. To do so, I discretize publisher slant into quintiles and calculate the first-order Wasserstein distance of engagement or promotion shares across these five bins of publisher slant relative to a uniform distribution. This distance metric is better suited to this setting relative to other common measures of diversity used in the literature, including the Herfindahl-Hirschman Index and Shannon Entropy, due to the ordered nature of slant partitions. For example, the Wasserstein distance between a user's engagement and the uniform distribution is larger (i.e. less diverse) for a user who engages only with publishers from a politically slanted partition versus a user only engaging with moderate publishers. The distance is minimized when users engage equally with publishers from all slant partitions and is largest when only engaging with publishers from a politically extreme partition.

The counterfactuals suggest that the random and non-personalized engagement maximizing algorithms lead to slight increases in engagement diversity relative to the actual algorithm. That said, the personalized algorithm results in a decline in engagement diversity, with the average Wasserstein distance of individual engagement shares relative to a uniform distribution increasing by 10.5%. This increase occurs for a large majority of users, with 71.6% of users experiencing a decline in their engagement diversity in the personalized engagement-maximizing counterfactual. To put this into perspective, under the actual ranking the maximum share of user engagement within a single slant partition averages 29.0%. This rises to 31.1% in the personalized engagement-maximizing algorithm, a relative increase of 7.3%. Turning to the credibility-maximizing algorithm, I also find a decrease in the diversity of engagement as the average distance to uniform engagement shares rose by 3.0% and the average user's share of engagement in their maximal publisher slant partition increased to 30.8%.

Impact on Discussion Quality

I now turn to studying the impact optimizing for engagement has on discussion quality, as measured through the sentiment of comments submitted to the platform. Table 1.5 demonstrates that both the non-personalized and personalized algorithms slightly elevate the share of negative-sentiment comments submitted by the average user relative to the existing algo-

Figure 1.5: Impact of Algorithm on Share of Engagement with High-Credibility Publishers



(a) Distribution of Change in High-Credibility Share

(b) High-Credibility Share by Baseline Credibility

Note: (a) Plots the empirical CDF of the change in the share of engagement with high credibility publishers under the counterfactual algorithms relative to the existing algorithm. (b) Plots binned mean credibility shares under the counterfactual algorithm against credibility shares under the existing algorithm. Regression line is a fourth-order polynomial fit.

rithm. Recall a negative-sentiment comment is significantly more likely to contain strongly negative emotions such as anger and disgust and more likely to be classified as toxic. Moreover, inspecting negative comments reveals they are often extremely vulgar and unlikely to contribute to the discussion in a meaningful way.

While the effect on the sentiment of the average user is small, personalization increases the variance in sentiment leading to some users commenting more positively while others are shown content that makes them respond negatively (Figure A.20a). Figure A.20b plots the relationship between the change in the negative-sentiment share of users against user preferences for publisher credibility and I find that users who prefer less-credible publishers have a larger increase in their negative-sentiment share. The same is true of users who prefer left-leaning outlets, consistent with the sentiment preference estimates in Figure A.18.

Impact on Publishers

Thus far, I have focused on the impact that different ranking algorithms have on users and the types of publishers with which they engage. Here, I change the unit of analysis to the publisher and summarize how the counterfactual ranking algorithms impact different types of publishers.

Figure A.21 plots the change in publisher market share by publisher slant (Figure A.21a) and publisher credibility (Figure A.21b).[19] Optimizing solely for engagement leads to a reallocation of market share from left-leaning publishers to right-leaning publishers and a slight increase in the market shares of low-credibility publishers. Optimizing for engagement with high-credibility publishers leads to a reallocation of engagement from politically slanted publishers to more neutral publishers and a reallocation from low- to high-credibility publishers.

Engagement-Credibility Trade-Off

The results thus far have compared engagement-maximizing algorithms with a credibility-maximizing algorithm. That said, platforms or society may balance these competing objectives in a more nuanced manner rather than preferring either extreme. I now describe the frontier of possible outcomes as λ, the weight placed on engagement with high-credibility publishers, is varied. Figure 1.6 plots this trade-off along with points corresponding to the total engagement-maximizing algorithm, credibility-maximizing algorithm, non-personalized engagement-maximizing algorithm, and non-personalized credibility-maximizing algorithm.

[19] Here, publisher market share is defined as a publisher's share of total engagement in the counterfactuals. This differs from how publisher market share would traditionally be defined, and one should think of market share in this context as the share of traffic from the platform.

As can be seen, moving to the credibility maximizing algorithm reduces engagement by 5.0%. Nevertheless, platforms can achieve over half of the increase in news diet quality from the credibility-maximizing algorithm for a 1.9% decrease in engagement. This change in engagement is similar in magnitude to the difference between the non-personalized engagement-maximizing algorithm and the personalized engagement-maximizing algorithm. However, the non-personalized algorithm does not meaningfully improve the quality of users' news diets, while the credibility-aware algorithm increases the average share of engagement with high-credibility publishers by 5.9 percentage points for approximately the same total quantity of engagement.

The shape of this frontier is also important, as the gradient is relatively flat around the engagement maximizing algorithms. This suggests that, for small decreases in engagement, the platform can drastically increase the share of engagement with high-quality publishers. However, this also means that small differences in preferences between the platform and society can lead to large discrepancies in outcomes along the credibility dimension – again highlighting the importance of aligning the ranking algorithm's objective function.

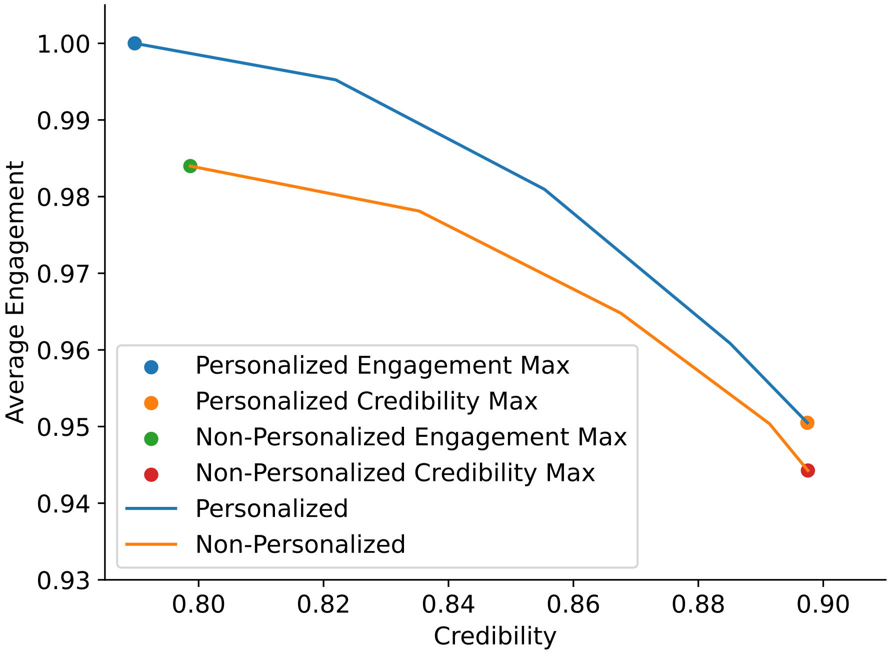
1.5.4 Robustness of Findings

Robustness to Alternative Method of Personalization and External Validity

The results in this section are based on a micro-founded model of user engagement decisions. This approach requires many assumptions and here I seek to show the results are not sensitive to the assumptions made. To do so, I summarize the findings of an analysis reported in Appendix A.4 that analyzes the type of publishers that are promoted when personalizing the ranking algorithm to maximize engagement using a reduced form collaborative-filtering based recommender system. The recommender system then recommends publishers on which a user is most likely to comment in a period. I validate this recommender system by estimating heterogeneous treatment effects when the regression discontinuity experiments align with the recommender system's predictions. I find that the recommender system effectively predicts treatment effects, a result that suggests the model has learned important aspects of user preferences. I then study the types of content that gets promoted under this simple recommender system to understand the extent to which personalized engagement maximization impacts individual news diets.

This approach has two advantages over the discrete choice model and counterfactual analysis I study in Section 1.4 and Section 1.5. First, this model is trained using comment decisions from over 500,000 users and is evaluated on comment decisions of over 180,000 users. This is a much larger sample than that used in the choice model approach, as I

Figure 1.6: Engagement-Quality Frontier



Note: This figure plots the frontier of possible outcomes when varying λ in the credibility aware algorithm. The y axis is average total engagement and the x axis is the average share of engagement with high credibility publishers. The y-axis is normalized to 1 at its maximal value. Points indicate outcomes under the counterfactual algorithms described in Section 1.5.2.

can use comment decisions on articles during periods not captured in Wayback Machine snapshots during the training process. Second, this approach relies on a different set of assumptions than the model of engagement decisions and can be validated by predicting treatment effects of which the model is predictive. I find consistent results across both approaches when comparable, which gives confidence that the findings of the choice model approach can be generalized to a broader set of users. That said, I emphasize the model-based approach as the main findings given it allows me to study in more detail the implications of various algorithm designs, including the credibility-aware ranking algorithms, on engagement rather than simply studying what content the algorithm would have recommended.

Robustness to Endogenous Search

A limitation of the analysis of counterfactual ranking algorithms presented above is the partial equilibrium nature of the analysis. That is, in equilibrium many things might adjust including user attention, the supply of articles, and the slate of existing comments on each article. Here we demonstrate the findings are not sensitive to endogenous attention allocation by microfounding the reduced form search process and allowing for endogenous search.

Recall in Section 1.4.1, a user is exposed to an article if $v_{ijt} = 1$ where v_{ijt} is an independent Bernoulli draw with probability $p\left(r,t\right)$. Here, we allow for a more flexible model of exposure where we model v_{ijt} as the composite of two random variables

$$v_{ijt} = v_{it} 1 \left[\bar{u}_{ir_j} > c_{ir_j} + \eta_{ijt}\right]$$

where v_{it} is an independent Bernoulli draw equal to one if user i is active in period t, $\bar{u}_{ir} = E\left[\max\left\{u_{ijt}, u_{i0t}\right\}|r_j = r\right]$ is the expected utility from viewing an article in position r, c_{ir} is the mean search cost of viewing an article in position r, and η_{ijt} is an idiosyncratic search cost. I assume that $\eta_{ijt} \sim N\left(0,1\right)$ and $\eta_{ijt} \perp v_{it}, x_{jt}, \xi_{jt}, \varepsilon_{ijt}$. I further assume for simplicity that search costs are such that there is no heterogeneity in user exposure probabilities (i.e. $c_{ir} = \bar{u}_{ir} - \Phi^{-1}\left(p_r\right)$). In equilibrium, this model of exposure is equivalent to the model in Section 1.4.1 but it allows for users to reoptimize p_r – the probability of being exposed to an article in position r conditional on being active – in the new equilibrium .

Under this model of exposure, the treatment effect of rank on engagement could come from two channels. First, articles in different positions have different search costs (i.e. $c_{ir} \neq c_{ir+1}$). For example, one may expect it to be more costly for users to view articles ranked lower on the page. Second, an article being promoted from position $r+1$ to position r would induce an update in user beliefs about the expected utility from viewing that article (i..e $\bar{u}_{ir} \neq \bar{u}_{ir+1}$). Under a counterfactual ranking algorithm, only this latter channel would change in equilibrium as users rationally update their beliefs about the expected utility they would receive from viewing articles in different positions. Therefore, understanding which channel drives the treatment effects we observe is important for understanding how attention may adjust in equilibrium.

Figure 1.7 plots the components of the search model with panel 1.7a showing the average $\bar{u}_{ir}$ by article position and counterfactual ranking algorithm and panel 1.7b showing the average c_{ir} by article position. As expected, both channels contribute to the estimated treatment effects with the average $\bar{u}_{ir}$ declining with rank and the average search cost increasing with rank. Moreover, for the various counterfactual ranking algorithms we also observe changes to $\bar{u}_{ir}$ as one would expect. The personalized engagement maximizing algorithm by construction has the sharpest decline in $\bar{u}_{ir}$ given the algorithm is explicitly promoting articles with high utility and therefore users update their beliefs about the quality of article they encounter. Also of note is that the credibility maximizing algorithm is non-monotonic. This is because a subset of users prefer low-credibility publishers and these are moved to the bottom of the page by this algorithm. Therefore, the expected quality of articles in these positions is higher for these users.

However, the magnitudes of the changes in $\bar{u}_{ir}$ are small relative to the search costs (Figure 1.7b). Therefore, the search costs are the primary driver of the treatment effects and allowing users to update reoptimize their attention across the feed has little impact on the counterfactual results. Figure A.23 plots the average $p_{ir} = P\left(\bar{u}_{ir_j} > c_{ir_j} + \eta_{ijt}\right)$ by rank and counterfactual algorithm and it is clear there is little change in the exposure probabilities and thus the main findings are robust to endogenous attention allocation within the feed.

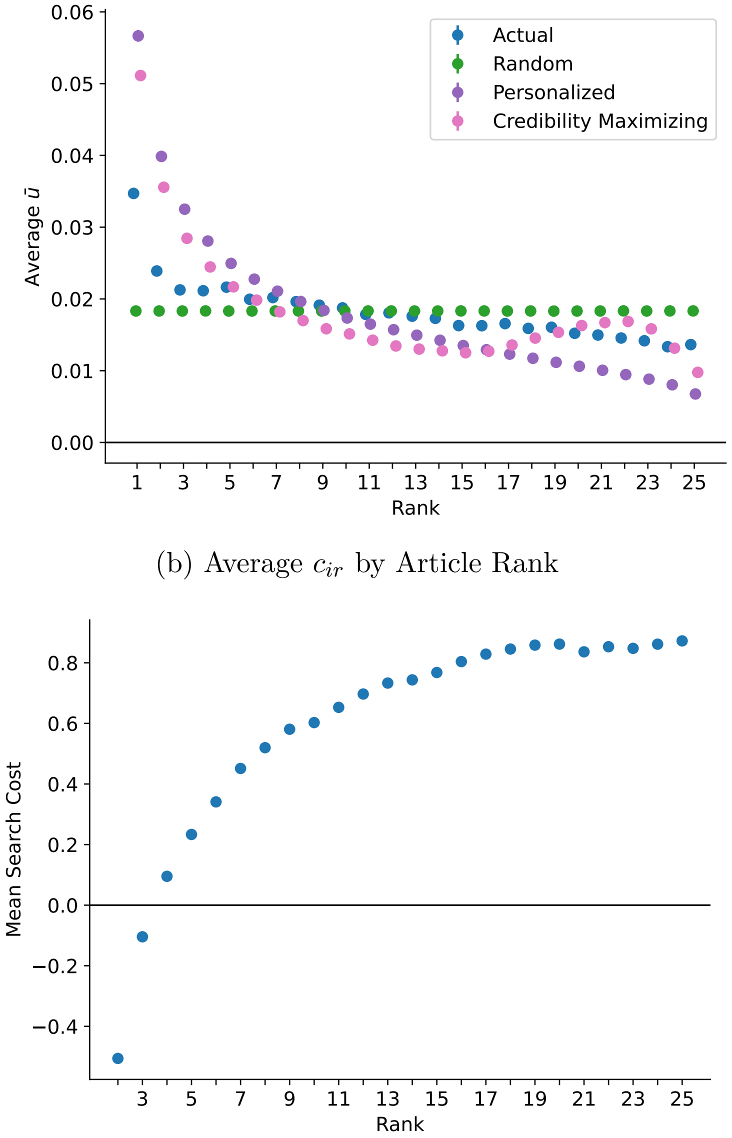
1.6 Discussion and Conclusion

In this study, I evaluate the impact of optimizing for engagement in social media news feed algorithms on the quantity, credibility, and diversity of publishers with which users engage. To address this question, I exploit a regression discontinuity design revealed in the platform's code to identify the causal effect of rank on engagement and use these causal estimates to identify a model of user engagement. Using this model, I estimate engagement patterns under counterfactual ranking algorithms including personalized and non-personalized engagement-maximizing and a credibility-aware algorithm that explicitly trades-off total engagement and engagement with high-quality publishers.

The counterfactual analysis demonstrates that social media platforms have a strong incentive to optimize their ranking algorithms for engagement. Optimizing for engagement leads to a dramatic increase in the quantity of engagement and much of this results from

Figure 1.7: Decomposing Treatment Effects in Endogenous Search Model

(a) Average $\bar{u}_{ir}$ by Article Rank



(b) Average c_{ir} by Article Rank

Note: (a) Average expected utility from viewing an article by position ($\bar{u}_{ir}$) and counterfactual algorithm. (b) Average search cost by article position (c_{ir}).

promoting posts with which all users are likely to engage. The marginal benefit of personalizing feeds is modest in terms of engagement quantity but has substantial impacts on the credibility and diversity of publishers with which users engage. In particular, personalized engagement maximization exacerbates differences in the quality of user news diets. That is, the personalized engagement-maximizing ranking expands the difference in the share of high-credibility engagement between users engaging with lower-credibility publishers and those engaging with higher-credibility publishers under the existing algorithm. In addition, personalization nearly uniformly decreases the diversity of publishers with which users engage.

Advertiser concerns about brand safety give platform managers a direct motive to promote credible publishers. Many advertisers seek to avoid advertising on platforms that promote content that is inconsistent with their values or that would create backlash from their consumers. For example, the #StopHateForProfit movement led over 1,000 large advertisers to halt or reduce advertising on Facebook to pressure the platform to expand its efforts to combat hate speech and misinformation [Hsu and Friedman, 2020, Hsu and Lutz, 2020]. There is also evidence suggesting that firms advertising on platforms alongside misinformation often experience customer backlash [Ahmad et al., 2023]. The credibility-aware algorithm demonstrates one method managers can use to improve the credibility of news content that is promoted on their platforms. The gradient of the engagement-credibility frontier indicates that moving away from the engagement-maximizing algorithm and towards the credibility-maximizing algorithm incurs a relatively small cost in terms of lost engagement. However, this also implies that small differences in the preferences of the platform and society can generate large changes in the amount of engagement with low-credibility publishers despite reasonably small changes to total engagement.

In addition, these results are also relevant for managers of publishers and the incentives they face when advertising revenue on traffic originating from social media referrals comprises an important component of their income. I find that personalized engagement maximization benefits publishers with a strong conservative slant and those producing low-quality journalism. This introduces an incentive for publishers to change their coverage to match the increased demand for politically slanted and low-credibility journalism.

Finally, these results have implications for regulating digital platforms. A growing regulatory trend is to require or incentivize platforms to allow users to opt-out of personalized recommendations or feeds. Examples include the European Union's Digital Services Act or proposed legislation (such as the Filter Bubble Transparency Act, Justice Against Malicious Algorithms Act, and the Protecting Americans from Dangerous Algorithms Act) in the United States. The findings presented here suggest the emphasis on allowing users to opt out of personalization may be misguided. Rather, as the results show, personalization has substantial benefits when the objective function aligns with social preferences. Recall that for approximately the same level of engagement as the non-personalized engagement-maximizing algorithm, the credibility-aware algorithm can increase the share of the average user's engagement with high-credibility publishers by 5.9 percentage points. To the extent that the platform's objective function differs from user preferences or those of society, a more efficient path forward would allow users to adjust the ranking objective function to align with their preferences or incentivize platforms to place the socially optimal weight on credibility.

Chapter 2

News Feeds and User Engagement: Evidence from the Reddit News Tab

2.1 Introduction

Social media platforms play an important role in everyday life, with the average person with access to the internet spending over two hours daily on social media [Kemp, 2020]. The prevalence of social media in modern life has important implications for individual well-being [Allcott et al., 2020], the quality of news people consume [Vosoughi et al., 2018b], and political polarization [Levy, 2021]. Perhaps the most common way users interact with many of the most popular social media platforms is through a news feed that aggregates content from across the platform (e.g. Facebook News Feed and Twitter timeline). A growing body of research has demonstrated the importance of news feed algorithms on individual well-being [Kramer et al., 2014], platform engagement [Dujeancourt et al., 2021], and exposure to counter-attitudinal news [Bakshy et al., 2015b, Huszár et al., 2021, Levy, 2021]. This paper studies the effects of a large social media platform introducing a popularity-driven news feed on engagement quantity, quality, and diversity. We find that the introduction of the feed increases engagement in some, but not all, featured communities and the engagement induced by the feed is heterogeneous in quality, with many communities seeing a larger relative increase in low-quality engagement compared to high-quality engagement. We also find suggestive evidence that the communities that saw the largest increases in engagement were communities most often featured in the new feed. Moreover, the introduction of the feed increases individual engagement diversity, suggesting non-personalized news feeds can be an important tool in mitigating algorithmic filter bubbles [Pariser, 2011].

This study leverages a natural experiment on the Reddit platform where a News tab offering popular content from a curated list of communities was introduced on iOS devices but not desktop or Android devices. At the time, Reddit statements indicated the iOS only roll out was to test and improve the feed and the platform had plans to introduce this feature across all devices. Therefore, we view this as an exogenous change to the app and argue that, absent the introduction of the News tab, Reddit engagement trends would have been similar across Android and iOS users. This allows us to identify the causal effect of the News Tab using a difference-in-differences strategy.

We find that the introduction of the news feed induces a statistically significant increase in the probability of posting any content on a news-related community. This increase is concentrated in the Politics, Technology, Entertainment, and Business related communities that were featured in the News tab. The News tab caused the monthly probability of posting by iOS users to increase by 19.6% (0.69 percentage points) in the Politics community, 17.3% (0.50 percentage points) in Technology communities, 7.0% (0.35 percentage points) in Entertainment communities, and 106.0% (0.21 percentage points) in Business communities. However, this new engagement that was caused by the News Tab is of heterogeneous quality. We find a statistically significant and economically meaningful increase in low-quality engagement as measured by voting on the platform in the Politics (36.0%, 0.23 percentage points) and Technology (13.1%, 0.08 percentage points) communities, and a non-statistically significant decline in Science and US & World communities. While the probability of posting a high quality comment also increased in the Politics and Technology communities, the increase was smaller in relative terms. Notably, the non-personalized feed also increases individual engagement diversity as implemented through the Shannon Entropy [Holtz et al., 2020, Shannon, 1948] and Herfindahl–Hirschman Index [Claussen et al., 2019, Rhoades, 1993]. This is true both for the diversity of engagement across the communities included in the News tab as well as for the diversity of engagement across articles from publishers of different political slant.

These results have important managerial and policy implications. In particular, we highlight the effects of social media news feed algorithms on engagement quantity, quality, and diversity. Despite the increase in engagement, the relative rise in low-quality engagement has the potential to make the platform less valuable to existing members by increasing the costs of finding high quality discussion and information [Gu et al., 2007]. This presents a trade-off, as prior work suggests the increase in engagement diversity caused by the News tab may have positive implications for user retention [Anderson et al., 2020, Oestreicher-Singer and Zalmanson, 2013]. While the increase in engagement diversity across categories of news has important managerial implications, from a policy perspective, the increase in the diversity of engagement across publishers from various political viewpoints suggests that non-personalized feeds can be an important tool to mitigate algorithmic filter bubbles.

2.2 Related Literature

In this paper we contribute to several streams of existing literature. First, we add to the literature studying the impacts of social media feed algorithms on users and society. This includes work studying the impact of social media feeds on individual well-being [Allcott et al., 2020, Kramer et al., 2014], media consumption [Allcott et al., 2020, Bakshy et al., 2015b, Levy, 2021], exposure to content from politicians [Huszár et al., 2021], and user engagement [Dujeancourt et al., 2021]. Kramer et al. [2014] find that changes to the Facebook News Feed that promote (or suppress) posts containing positive expressions cause users to post more (less) positive posts and less (more) negative posts, highlighting the importance of the News Feed algorithm in determining the content that users interact with and downstream effects on user behavior. Similarly, Bakshy et al. [2015b], Allcott et al. [2020], and Levy [2021] find that the news feed impacts the news people read, and in particular, exposure to counter-attitudinal sources. We contribute to this literature by further demonstrating the importance of news feed design on user behavior. In particular, the addition of the news feed we consider increases user engagement, and the diversity of user engagement, suggesting that non-personalized feeds can help mitigate filter bubbles that are often generated by social media feeds.

Second, we contribute to the growing literature studying the impacts of news feeds and algorithmic recommendations on consumer behavior. This has been studied in the context of product sales [Hosanagar et al., 2014, Lee and Hosanagar, 2019, Oestreicher-Singer and Sundararajan, 2012] as well as content consumption [Bakshy et al., 2015b, Claussen et al., 2019, Dujeancourt et al., 2021, Holtz et al., 2020]. The existing work find that personalized algorithms increase content consumption relative to manually curated recommendations [Claussen et al., 2019, Holtz et al., 2020] and chronologically ordered news feeds [Dujeancourt et al., 2021]. In addition, Bakshy et al. [2015b], Claussen et al. [2019] and Holtz et al. [2020] find (to varying extents) that personalized recommendations decrease individual consumption diversity, supporting the notion of algorithmic recommendations leading to filter bubbles. We contribute to this body of work by studying the causal effects of a new non-personalized feed on engagement and engagement diversity. In contrast to the work studying personalized feeds [Bakshy et al., 2015b, Claussen et al., 2019, Holtz et al., 2020], we find the non-personalized news feed increases individual engagement diversity and this increase in diversity does not come at the expense of decreased engagement. Moreover, this paper is among the first to study the impacts of a new feed at a major social media outlet on engagement quantity, quality, and diversity.

Finally, we add to the literature studying motivations for user generated content. Past research has investigated numerous interventions to induce additional user generated content. These interventions include financial incentives with mixed results [Burtch et al., 2018, Cabral and Li, 2015, Khern-am nuai et al., 2018], successful social norm interventions [Burtch et al., 2018, Chen et al., 2010], and status or rewards [Burtch et al., 2021, Gallus, 2017, Goes et al., 2016, Restivo and Van De Rijt, 2012]. Of particular relevance to this work are the trade-offs faced in stimulating additional user generated content. For example, Khern-am nuai et al. [2018] find that financial incentives can increase the quantity of online reviews, though these incentives result in the marginal reviews being of lower quality. Gu et al. [2007] emphasize this tradeoff explicitly and study the competing positive network externalities, stemming from additional engagement providing more information, and negative externalities, if additional low-quality engagement distracts members of the community and increases costs of finding relevant information. We contribute to this literature by asking if additional news feeds can stimulate user-generated content, in addition to the impacts these have on the quality of this engagement.

2.3 Setting

Reddit is a popular social media platform founded in 2005 with over 52 million daily active users as of January 2020.[1] The platform consists of over 100,000 active communities called subreddits, which host user-generated content focused on a particular topic. Within a community, users can post new submissions or comment on others' submissions. By default, content is presented to users using a proprietary algorithm that favors upvotes and fresher content.[2]

Voting on content is an important part of Reddit both practically, as it is a key driver of content promotion, and as a method of rewarding content that contribute to the community and demoting those that do not.[3] While norms vary within communities, Reddit guidelines are explicit that voting should reflect contributions to the community and conversation.[4] In particular, downvoting only because you disagree with the content is explicitly discouraged and downvoting should be reserved for content that is not contributing to the community's conversation. Therefore, in this study we will use voting data to infer post quality as judged

[1] https://www.redditinc.com/press

[2] The exact details of this algorithm were publicly available until 2017 but are now proprietary.

[3] Upvoting (downvoting) a user's post or comment impacts their Karma score, which is a publicly available number summarizing "how much good the user has done for the reddit community" (https://www.reddit.com/wiki/faq).

[4] "Vote. If you think something contributes to conversation, upvote it. If you think it does not contribute to the subreddit it is posted in or is off-topic in a particular community, downvote it." (https://www.reddithelp.com/hc/en-us/articles/205926439)

by members of the community.

Reddit is accessible to users through web browsers or mobile apps. In April 2016 Reddit launched their official mobile apps for Android and iOS devices. Before the official Reddit apps were supported, there were a number of third party apps that allowed users to browse the site and many unofficial apps are still available today, though the official Reddit app is the dominant app in the market.[5] Users of the official mobile app are able to access three primary sections of the app through a navigation bar at the top of the screen: "Popular", "Home", and "News", though the latter News section is only available to users with iOS devices. The Popular tab aggregates popular content from across the site and the Home tab aggregates content from communities in which the user is a member. The News tab is the focus of this study and is discussed in detail in Section 2.3.1.

2.3.1 Natural Experiment

In June 2018, Reddit introduced an update to its mobile app on Apple (iOS) devices that introduced the News tab, which provided a feed of content from communities that focus on discussion and sharing of news related content. The tab is displayed prominently in the mobile app alongside the Home tab that shows submissions from communities a user is a member of and the Popular tab that shows popular content from across the platform (Figure 2.1). Within the News tab, users first view a feed containing posts from all news categories. Users may then select individual topics to view more focused feeds that display posts related to the selected topic. The communities that are displayed in the News tab are chosen to be those that most often engage with news, who are actively moderated and in compliance with Reddit policies on acceptable content and guidelines for healthy communities, and who require the title of posts linking to news articles to be an accurate reflection of the article title. These guidelines result in most posts in the News tab following a common structure, where the post title is an article headline and the body links to the full article (Figure 2.1).

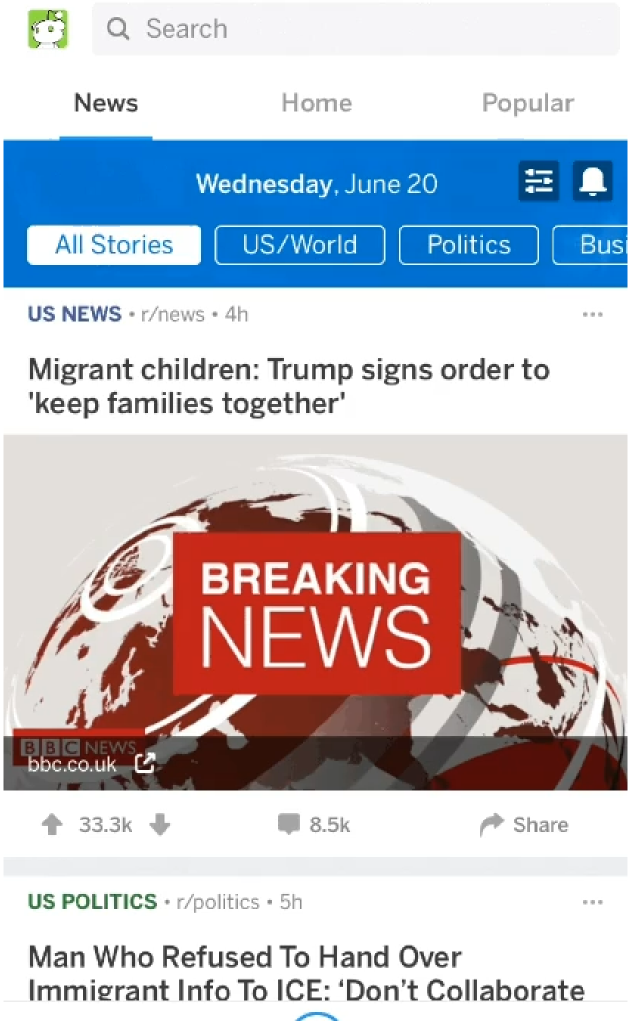
In Reddit's public comments at the time, they announced that the News tab was originally being released on iOS, but would eventually be available on most devices.[6] Our empirical strategy, which will be discussed in greater detail in Section 2.4.3, relies on the assumption that, absent the introduction of the News tab, engagement trends of iOS users in our sample would have followed engagement trends of Android users in our sample.[7] We

[5] While data on installs is not publicly available, the official Reddit applications have received more than 2 million reviews on both the Apple App Store and the Google Play Store. The next closest competitor has roughly 400,000 reviews.

[6] https://www.reddit.com/r/announcements/comments/8sth30/extra_extra_were_launching_a_news_tab_as_a_beta

[7] Our preferred results require a slightly weaker assumption that common trends hold conditional on

Figure 2.1: Screenshot of Reddit News Tab



US NEWS • r/news • 4h

Migrant children: Trump signs order to 'keep families together'

US POLITICS • r/politics • 5h

Man Who Refused To Hand Over Immigrant Info To ICE: 'Don't Collaborate provide evidence consistent with such an assumption in Section 2.4.4, though the assumption cannot be explicitly tested empirically.

2.4 Data

The data for this study are based on a dataset of public Reddit submissions and comments described in Baumgartner et al. [2020].[8] We focus on posts between June 2017 and June 2019 which contain a total of 349 million submissions and 3.0 billion comments during this period. We restrict our sample to the subset of users for whom we can infer their mobile device, and this sample has 28.0 million total comments across all communities during the period. We also focus on comments rather than submissions, as comments make up the majority of posts and this is particularly evident in communities promoted on the News tab. In our sample of users, comments on communities featured in the News tab account for 97.89% of all posts. Table 2.1 below shows descriptive statistics for our sample.

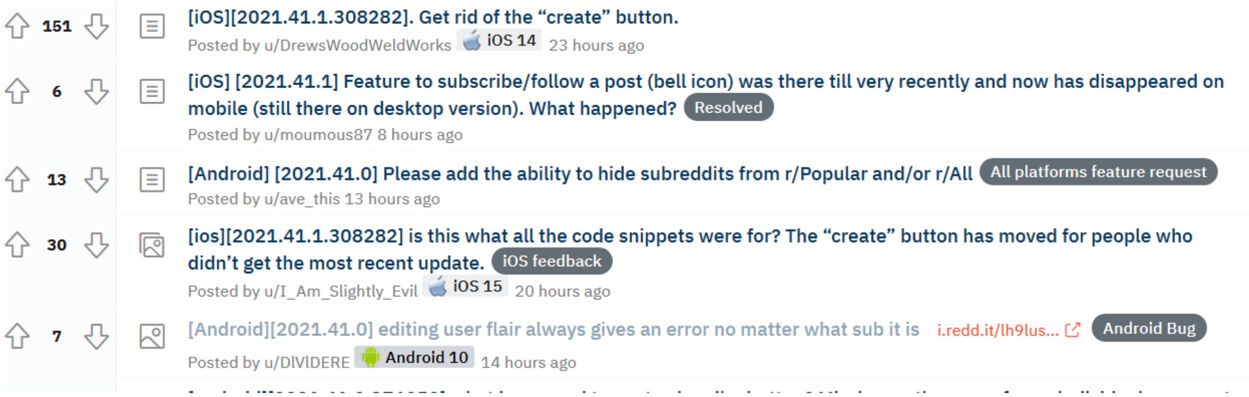
observed pre-treatment engagement. This is because we use Coarsened Exact Matching weights in the analysis [Iacus et al., 2012], described in Appendix B.1.

[8] Gaffney and Matias [2018] find evidence of missing data in this dataset. In particular, they find that less than 0.04% of comments and 0.65% of submissions are missing from the dataset in the early years. We believe missing data on this scale is unlikely to be driving our results for two reasons. First, as discussed in Baumgartner et al. [2020], the data collection process has improved as a result of the flaws highlighted in Gaffney and Matias [2018] which analyze data before the period studied here. Second, Gaffney and Matias [2018] find that heavy users are more likely to be impacted than light users. Our primary outcomes are indicators if a user posted *any* posts in a month. Therefore, for the outcome to be changed we would have to be missing all of a users posts in a given month which is less likely.

Table 2.1: Summary Statistics

	Any Post	Number of Posts			Any Neg Score		Any Pos Score	
	Mean	Mean	Min	Max	Mean	P(any|posting)	Mean	P(any|posting)
All Posts	0.625	541.057	0	91,681	0.464	0.742	0.624	0.998
All News	0.374	40.856	0	12,927	0.166	0.444	0.369	0.985
US/World	0.213	5.641	0	4,298	0.074	0.349	0.206	0.969
Politics	0.106	6.277	0	9,369	0.036	0.338	0.102	0.960
Technology	0.142	1.444	0	2,253	0.031	0.217	0.137	0.965
Science	0.117	1.128	0	2,182	0.019	0.160	0.113	0.965
Sports	0.131	16.624	0	12,740	0.045	0.346	0.127	0.973
Gaming	0.178	6.389	0	4,289	0.051	0.287	0.173	0.976
Entertainment	0.172	3.227	0	3,063	0.042	0.242	0.167	0.969
Business	0.014	0.126	0	264	0.003	0.181	0.014	0.951

Figure 2.2: Screenshot of Reddit Mobile Feed Used to Identify User Devices



151 [iOS][2021.41.1.308282]. Get rid of the "create" button.
Posted by u/DrewsWoodWeldWorks iOS 14 23 hours ago

6 [iOS] [2021.41.1] Feature to subscribe/follow a post (bell icon) was there till very recently and now has disappeared on mobile (still there on desktop version). What happened? Resolved
Posted by u/moumous87 8 hours ago

13 [Android] [2021.41.0] Please add the ability to hide subreddits from r/Popular and/or r/All All platforms feature request
Posted by u/ave_this 13 hours ago

30 [ios][2021.41.1.308282] is this what all the code snippets were for? The "create" button has moved for people who didn't get the most recent update. iOS feedback
Posted by u/I_Am_Slightly_Evil iOS 15 20 hours ago

7 [Android][2021.41.0] editing user flair always gives an error no matter what sub it is i.redd.it/lh9lus... Android Bug
Posted by u/DlVlDERE Android 10 14 hours ago

2.4.1 Inferring Device Type

A drawback of the Baumgartner et al. [2020] data relative to the proprietary data collected by the platform is that we only observe publicly available information, which does not include the device a user was using when making a post. As a result, we must infer device type from posts on the platform. To do so, we consider the subset of users who have posted in the RedditMobile community, which is an official Reddit community for announcements, discussion, and feedback on the official Reddit mobile apps. When posting in this community, users are often posting feedback about the mobile apps and typically include explicit tags about the device and version of the mobile app they are giving feedback (Figure2.2). We infer user device using the following procedure. First, if the user tagged a particular operating system in their post we assign their device accordingly. Second, if the user has tagged their operating system on the community through author 'flair' (for example, the first post in Figure 2.2 has tagged iOS 14 as their author flair), we assign them to that device. Finally, if both of these methods fail and the post is a comment, we assign the commenter the device of the post they are commenting on. This procedure allowed us to identify 18,274 Android users and 19,127 iOS users. There were an additional 1,579 users who authored posts that would have been classified as both Android and iOS devices, and these users are excluded from all analyses.

The process of inferring device operating systems limits our sample substantially. As a result, we include all users who have posted on the RedditMobile community rather than only those posting before the News tab was introduced. The primary risk of inferring device type from posts after the News tab was introduced is a bias if users select devices based on this intervention, but we believe this is unlikely to be driving our results.[9] A more likely

[9] When the News tab was announced, Reddit administrators were explicit that they intended to make this feature available across all devices which should mitigate switching im-

issue is device switching that is independent of the Reddit News tab. This should bias our estimates of engagement toward zero as long as the News tab did not induce some users to engage less with news related communities.

2.4.2 Communities of Interest

The News tab includes 54 communities (subreddits) that are structured into 8 higher-level categories of news.[10] In this study we aggregate engagement to the category level and focus on heterogeneity along news categories rather than individual communities. We do this because the News tab is structured around these categories and the smaller number of categories facilitates comparisons of heterogeneous effects.

We exclude a handful of communities due to concerns about the common trends assumption. First, we exclude technology communities that specifically reference Apple or Google as we are conditioning on users having an Apple or Android device and expect these users to have different engagement patterns on these communities. Second, we exclude all communities in the Crypto category. This is because of the massive increase in traffic in late 2017 that resulted in differential engagement by iOS and Android users. In the end, we focus on the following 8 categories of news: US/World, Politics, Technology, Science, Sports, Business, Gaming, and Entertainment.

2.4.3 Empirical Strategy

We estimate the effect of the News tab on Reddit activity using a difference-in-differences design that makes a common trends assumptions. Formally, we model outcomes Y_{it} using a two-way fixed effects panel regression model

$$Y_{it} = \alpha_i + \lambda_t + \tau \text{Post}_t D_i + \varepsilon_{it} \tag{2.1}$$

where Y_{it} is the outcome of interest, D_i is an indicator equal to one if user i has an iOS device, Post_t is an indicator for the post-treatment period, and α_i (λ_t) represent unobserved

mediately following the release. The announcement referenced specific plans for availability on desktop and a top comment from an administrator stated the intention to make this available across all platforms (https://www.reddit.com/r/announcements/comments/8sth30/extra_extra_were_launching_a_news_tab_as_a_beta/e12auz7?utm_source=share&utm_medium=web2x&context=3). As it became evident that this feature was not planned to be released on Android, selective switching is more plausible, though we find it unlikely that a substantial share of our sample is choosing a smartphone operating system because of this particular Reddit feature.

[10] When Reddit launched the News tab in 2018, the list of communities that were referenced were not made public. To approximate the list of communities that are promoted by the News tab we consider the communities that are present as of August 2021.

individual (time) fixed effects. Identification of this model comes from a common trends assumption that assumes, absent the introduction of the News tab, the average outcome for iOS and Android users would have had the same variation over time [Abadie and Cattaneo, 2018].

In addition to average effects, we study dynamic treatment effects by estimating event-study models of the form

$$Y_{it} = \alpha_i + \lambda_t + \tau_t D_{it} + \varepsilon_{it}. \tag{2.2}$$

We then can interpret τ_t as the average treatment effect on iOS users in period t to understand how the treatment effect varies over time. In addition, this specification forms the basis for our test of pre-trends discussed in more detail in the following section. All statistical inference on estimates of Equation 2.1 and Equation 2.2 use cluster robust standard errors clustered at the user level [Liang and Zeger, 1986].

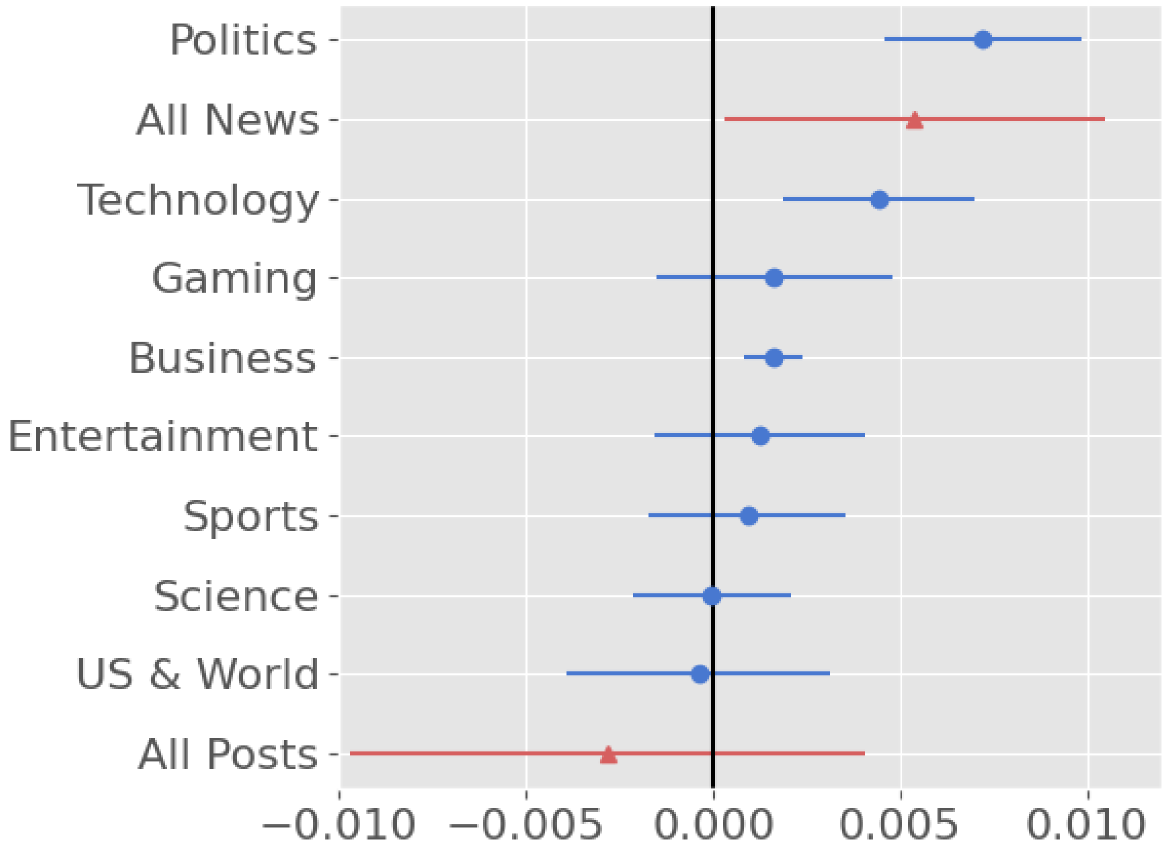
Before estimating Equation 2.1 and Equation 2.2 we perform a Coarsened Exact Matching (CEM) procedure that accounts for imbalance in baseline engagement in the pre-period through re-weighting [Gertler et al., 2016, Iacus et al., 2012]. The matching procedure is explained in detail in Appendix B.1. This weakens the identification assumption required, as we now only need common trends to hold conditional on the covariates used in matching. Results of the analysis without using the CEM weights are shown in Appendix B.7 and the results are largely consistent.

2.4.4 Testing for Pre-Trends

To have a causal interpretation, the empirical strategy described in Section 2.4.3 relies on a common trends assumption. To be explicit, our identifying assumption is that, absent the introduction of the News tab, iOS and Android engagement would have followed the same time trends conditional on pre-treatment covariates used in matching. While this cannot be tested empirically, we can test for common trends in the pre-treatment period that would be consistent with this assumption. To do so, we perform the joint test that pre-treatment coefficients in Equation 2.2 are equal to zero.

When considering if a user posted in any news related community, we fail to reject the null hypothesis of common pre-trends (p=0.36). Moreover, when looking at specific topics we fail to reject the null of common pre-trends at the 5% level in all cases except for entertainment related communities, which we can reject with a p-value of 0.05 (Figure B.2). For the outcome of an indicator of any low-quality post, we fail to reject the null hypotheses of common pre-trends for all 8 categories of news (Figure B.3) and the same is true for the high-quality post analysis (Figure B.4).

Figure 2.3: Treatment Effect Estimates on Probability of Posting a Comment



Note: Coefficients from estimates of Equation 2.1 on an indicator if a user posted in the community in a given month. Each point represents the estimated average treatment effect on iOS users on probability of engagement for the 8 categories of news. Bars represent 95% confidence intervals.

2.5 Results

2.5.1 Effect on Engagement Quantity

To study the impact of the News tab on engagement with news related communities, we first estimate Equation 2.1 where the outcome is an indicator equal to one if a user posts on any community suggested by the News tab. In aggregate, we find the News tab increases the probability of posting on any news related community by 3.5% (0.61 percentage points percentage points, p=<0.01).

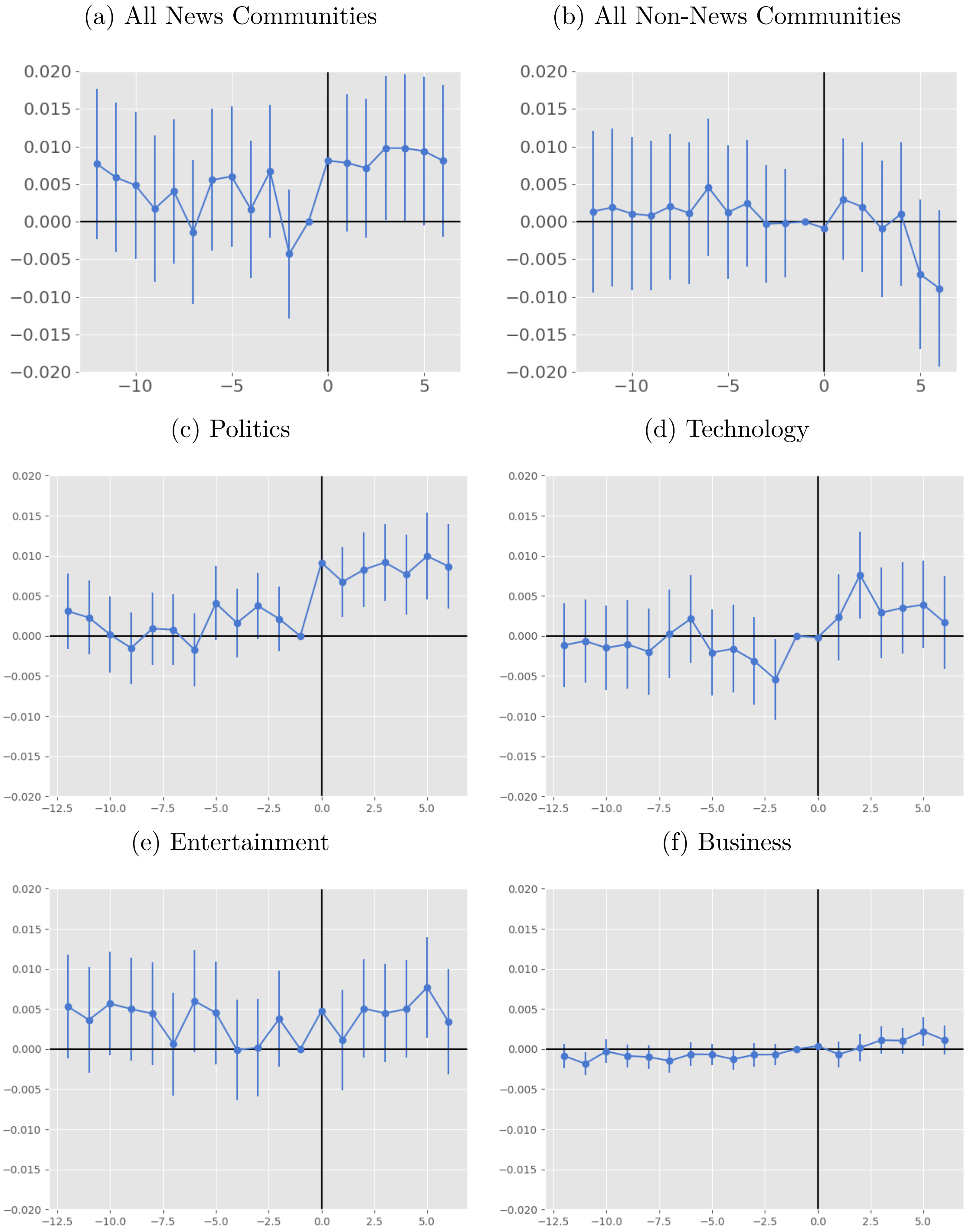
This aggregate view, however, masks substantial variation in effect sizes by the topic of news (Figure 2.3). Recall there are 8 categories of news included in the News tab (US/World, Politics, Technology, Business, Science, Sports, Gaming, and Entertainment) and we can estimate the effect of the News tab on the monthly engagement probability for each category. Figure 2.3 plots the estimates of the average treatment effect on iOS users of the News tab on engagement with each of the 8 news categories. In all categories, the treatment effect point estimates are positive, though some are statistically indistinguishable from 0. This suggests the News tab increases total engagement, but this increase is concentrated in a subset of news categories. There is a statistically significant treatment effect for the Politics (p=<0.001), Technology (p=<0.001), Entertainment (p=<0.01), and Business (p=<0.001) categories that are also significant when correcting for multiple hypothesis testing [Holm, 1979].

Recall these effect sizes represent the share of individuals induced by the News tab to post in each particular community in a given month. Baseline engagement rates are relatively low in this sample, with on average only 3.5% of iOS users posting in the Politics community in the year leading up to the introduction of the News tab. Therefore, a treatment effect of 0.69 percentage points represents a 19.6% increase in the *monthly* share of iOS users who post in the Politics community. There is also a 17.3% increase in the share posting in Technology communities, a 7.0% increase in the share posting in Entertainment communities, and a 106.0% increase in the share posting in Business communities.

The above analysis focuses on the extensive margin of engagement, showing that the News tab induces additional iOS users to post on news related communities relative to Android users. Next, we consider how the News tab impacts the intensive margin by estimating Equation 2.1 on a series of additional outcomes. Specifically, we estimate Equation 2.1 where the outcome is a series of indicators equal to one if total engagement in a category is above a threshold ranging from 0 to 50. This estimates the treatment effect for iOS users on the probability of a user having more posts in a month than the threshold.[11] The results of this analysis are shown in Figure B.7 and Figure B.8. There is a clear pattern, where the absolute treatment effects of the News tab are largest for lower thresholds suggesting the News tab induces new users to post a few times rather than inducing users to post more regularly.

[11] A natural outcome for the extensive margin analysis would be the log-transformed number of posts, though the log-transformation suffers two pitfalls. First, given the sparsity of our dataset the arbitrary choice of how to handle zeros would be consequential. Second, and more importantly, common trends in the extensive margin (probability of posting) is inconsistent with common trends in the log-transformed outcome unless we make further strong assumptions that are rejected in the data.

Figure 2.4: Dynamic Treatment Effect Estimates on Probability of Posting a Comment



(a) All News Communities

(b) All Non-News Communities

(c) Politics

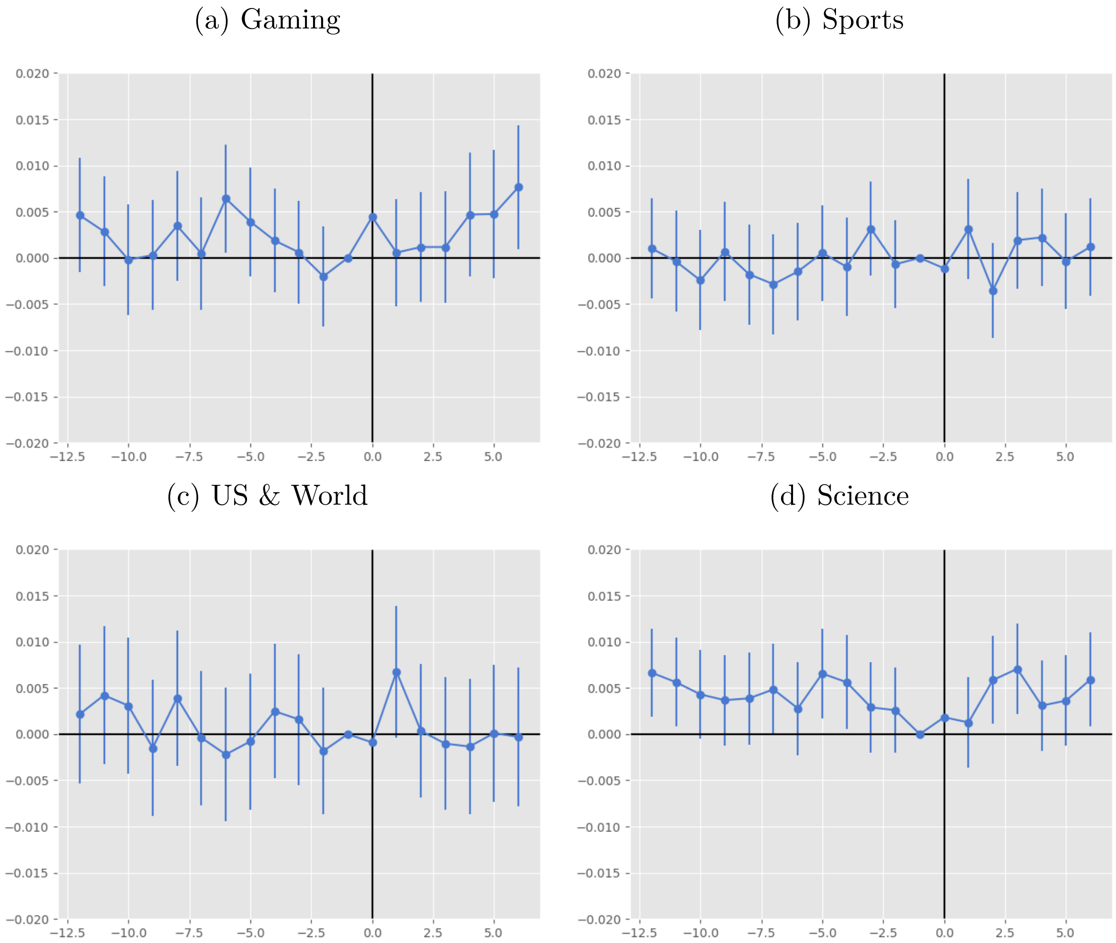
(d) Technology

(e) Entertainment

(f) Business

Note: Dynamic treatment effect estimates on probability of posting in community, estimated using Equation 2.2. Bars represent 95% confidence intervals.

Figure 2.5: Dynamic Treatment Effect Estimates on Probability of Posting a Comment


(a) Gaming

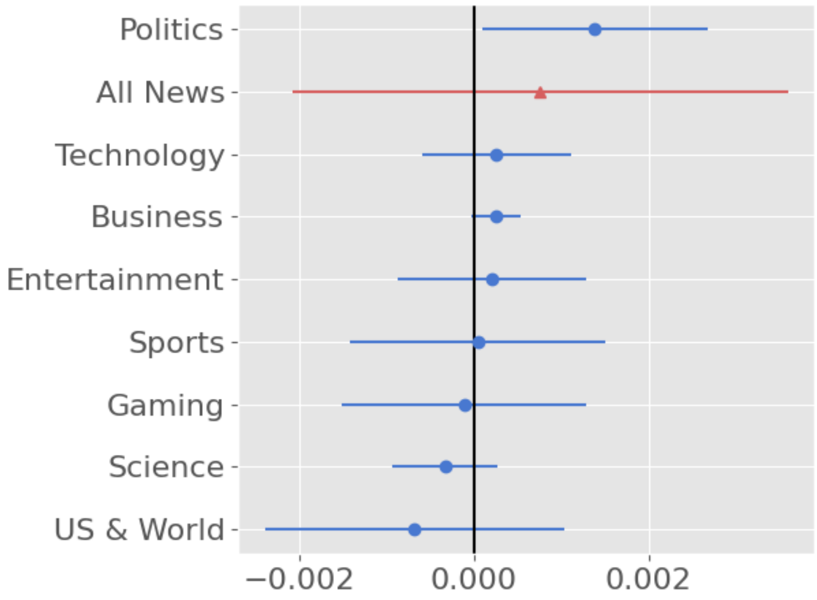
(b) Sports

(c) US & World

(d) Science

Note: Dynamic treatment effect estimates on probability of posting in community, estimated using Equation 2.2. Bars represent 95% confidence intervals.

Figure 2.6: Treatment Effect on Probability of a Post by Post quality



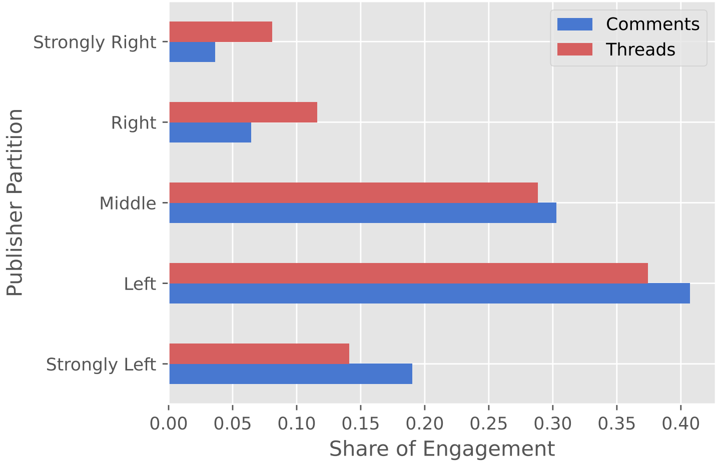
Note: Coefficients from estimates of Equation 2.1 on an indicator if a user posted a positive (negative) post in a given month. Each point represents the estimated average treatment effect on iOS users on the probability of posting a positive or negative comment for the 8 categories of news. Bars represent 95% confidence intervals.

2.5.2 Effect on Engagement Quality

In addition to the quantity of engagement, we also observe the quality of engagement from upvotes and downvotes by participants in the communities. Recall that votes on Reddit are intended to be a mechanism to signal if a comment or submission contributes to the conversation or not, which we are interpreting as a signal of post quality as judged by the participants in the community.

To study the impact of the News tab on engagement quality, we again estimate Equation 2.1 with an indicator if any of an author's posts were 'negative' or 'positive'. Here, we define a negative post to be a comment that received more downvotes than upvotes and a positive post to be a comment that received more upvotes than downvotes. Results of these regressions are plotted in Figure 2.6. We find that there is a heterogeneous effect on engagement quality by topic. For example, while the Politics and Technology communities see an increase in the number of users with both positive and negative posts, there is a relatively larger increase in the share of users with negative posts suggesting the marginal comments induced by the News tab are lower quality in these communities.

Figure 2.7: Distribution of Posts in Hard-News Communities Across Domain Slant Partitions



Note: For the set of hard news communities (US & World, Politics, Business), this figures plots the share of engagement on posts started by a publisher across the political spectrum.

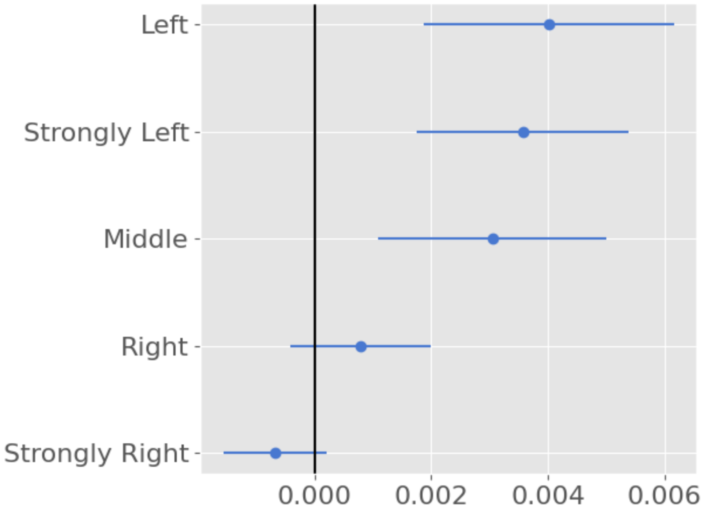
2.5.3 Effect by Domain

Focusing on news categories we classify as "hard news", which includes US & World, Politics, and Business, we now investigate the effect of the News tab on the political slant of the news publishers that users engage with. In particular, we consider heterogeneous effects on engagement with news articles by publishers of varying political slant. Appendix B.2 explains in detail how political slant is measured for each publisher.

The vast majority of threads in the news communities are started by someone sharing an article related to the community's topic. For this analysis, we drop the 4.2% of posts that link to Reddit (this is primarily general discussion threads) and YouTube (<0.1% of posts). We match publisher domains to the domain political slant measures of Robertson et al. [2018], who calculate domain level political slant of 19,022 of the most popular domains. Additional details about these data can be found in Appendix B.2. Over 94% of the remaining posts in hard news communities by our sample were on a thread started by a link to a publisher domain contained in the Robertson et al. [2018] slant data. We then partition the publishers into five equally sized bins based on their slant. The majority of engagement is on articles from left-leaning publishers, with less than 20% of posts on threads initiated by articles from right-leaning outlets and over half of posts on threads initiated by articles from left-leaning outlets (Figure 2.7).

Next we investigate how the News tab differentially impacts engagement on the various partitions. We find that the News tab increases engagement across the political spectrum, though point effects are largest among strongly left and moderate publishers. The increase

Figure 2.8: Treatment Effects on Probability of Posting on a Thread by Publisher Political Slant



Note: This figure plots estimates of Equation 2.1 where the outcome is an indicator if the user posts on a thread in a hard news community discussing an article from a publisher within each slant partition.

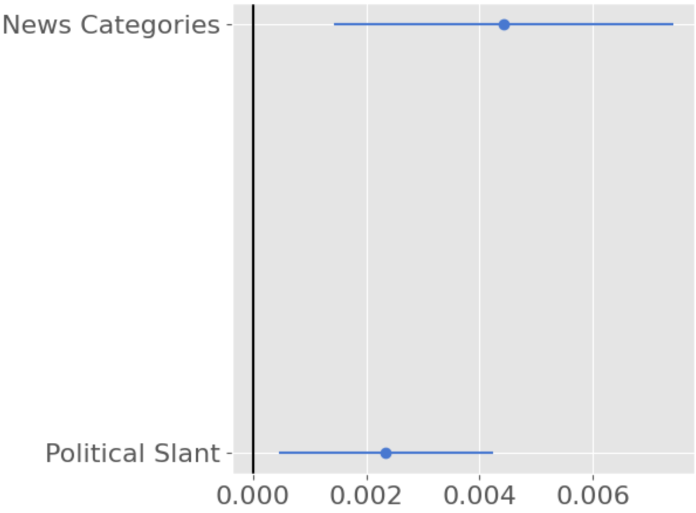
among conservative-leaning publishers are statistically indistinguishable from zero (Figure 2.8).

2.5.4 Effect on Individual Engagement Diversity

Turning toward the diversity of engagement, we find that the News tab induces individuals to engage with more diverse content. This is true both among the topics included in the News tab and among publisher slant partitions within "hard news" communities (Figure 2.9). Diversity here is operationalized as the Shannon Entropy [Holtz et al., 2020, Shannon, 1948] of engagement shares by news category and publisher slant partition, respectively. Full details of the diversity measure can be found in Appendix B.3 where we also demonstrate the robustness of this result to other diversity measures, including the Herfindahl-Hirschman Index.

We find that the Shannon Entropy of diversity across categories of news included in the News tab increased by 12.0% and the News tab increased the Shannon Entropy of diversity across publishers from different political slants by 8.2%. To try and contextualize the magnitude of this increase in diversity, we calculated what share of the maximum possible increase in diversity we observed from the News tab. Specifically, for iOS users in the post-treatment period, we calculated the maximum possible Shannon Entropy of engagement

Figure 2.9: Effect of News Tab on Engagement Diversity



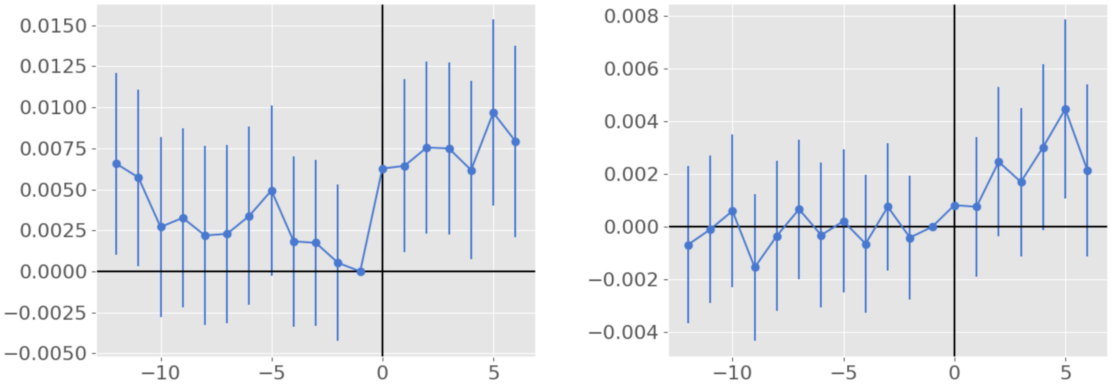
Note: Treatment effect estimates of Equation 2.1 of individual engagement diversity by device, as measured through the Shannon entropy. The News Category estimate reflects the treatment effect on diversity of engagement across the \ncats \ categories of communities on the news tab. The Political Slant estimate reflects the treatment effect on engagement diversity across political slant partitions within hard news communities.

given the total engagement amount and estimated Equation 2.1 on this new outcome. This provides an upper bound on the potential increase in engagement diversity holding total engagement fixed. We find that the increase in engagement caused by the News tab represent 3.8% of the maximum increase for diversity across news categories and 7.7% of the maximum increase for diversity across publishers of varying political slant.

2.6 Heterogeneity and Robustness

As shown in Section 2.5.1, there is substantial heterogeneity in the impact of the News tab on engagement with the various categories of news, with the Politics community seeing the largest increase in engagement. A potential explanation for this heterogeneity could be a result of the design of the News tab itself. While the News tab does have subsections for each of the 8 different categories of news content, the feed individuals first interact with aggregates popular content from across all categories of news content (Figure 2.1). It is plausible that users are more likely to engage with content promoted on this page. To evaluate this hypothesis, we create a monthly index of popularity for each news category

Figure 2.10: Individual Engagement Diversity Time Series



(a) Across Community Engagement Diversity (b) Political Slant Engagement Diversity

Note: These plot the estimates of Equation 2.2 of individual engagement diversity by device, as measured through the Shannon entropy. Figure 2.10a plots diversity of engagement across the 8 categories of communities on the news tab. Figure 2.10b plots engagement diversity across political slant partitions within hard news communities.

and predict monthly treatment effect estimates with this index.[12] We find that this measure of popularity is correlated with the monthly treatment effect estimates on the probability of posting. While this cannot be interpreted causally, as popularity is endogenous, this evidence is consistent with the hypothesis that the heterogeneous effects of the News tab are a result of the underlying popularity of the different categories. An important implication of this hypothesis is that the choice of the popularity algorithm is critical. If this hypothesis were correct, it would mean promoting content from a different category of news would increase engagement with that category, More important, it would also suggest an algorithm that favored content slanted towards a particular political party would increase engagement with this content (e.g. Huszár et al. [2021]), further highlighting the importance of the choice of algorithm in influencing the content individuals see and ultimately engage with.

A competing hypothesis that could explain why the largest treatment effects are in the Politics category is that the News tab was introduced in the months leading up to a U.S. midterm election. To address the concern that the our results are confounded by the election, we estimated the effect of the News tab on placebo communities that are focused on discussing politics but were not included in the News tab politics section. We find precise null effects for these communities suggesting that the effect we find for the Politics commu-

[12] Unfortunately, historical data on which posts were promoted are unavailable. Therefore, we scrape a Reddit page that shows the most popular posts from a given day. See https://www.reddit.com/r/changelog/comments/k663qy/introducing_rereddit_go_back_in_time_to_see_top/ for a description of this page. We then define the popularity index as the monthly average share of posts in the top 50 most popular news posts that came from each category of news.

Table 2.2: Correlation in Community Popularity and Treatment Effects

	(1)
Popularity	0.049
	(0.010)
Constant	0.003
	(0.000)
p-value	<0.001
Obs	56
Adj. R^2	0.232
F-stat	21.827

Note: Coefficients from the regression of monthly treatment effect estimate on community popularity. Model (2) adds community fixed effects. Model (3) adds community fixed effects and the interaction of centered community fixed effects with popularity. This results is equivalent to averaging the estimates.

nity included in the News tab is in fact a result of the News tab and not a result of the introduction of the News tab coinciding with an election cycle (Figure B.6).

2.7 Discussion and Conclusion

We find that the News tab increased engagement with news related content, though the marginal content induced by the News tab was of heterogeneous quality. In addition, the News tab increases the diversity of engagement across news categories and content from outlets with different political leanings.

This study, however, is not without limitations. First, our analysis relies on publicly observed information which requires us to limit our sample to individuals who reveal their device through the RedditMobile community. While we believe this does not impact the internal validity of our results, we must be cautious when generalizing these results as this sample of users may not be representative of the broader population. In addition, our data only contain engagement measured by posting to the communities and we do not observe content consumption.

Our findings have several important takeaways. First, we highlight the heterogeneous increase in engagement that was induced by the News tab. We found heterogeneity in two dimensions. Not all communities saw significant increases in engagement with only Politics, Technology, Entertainment, and Business related communities seeing significant increases. In addition, the increased engagement was of heterogeneous quality as some communities (namely, Politics), also saw a significant increase in the number of users posting low-quality content. This highlights an important trade-off in platform design. Existing work suggests design changes that increase engagement, in particular the diversity of engagement, in virtual communities may improve retention of the affected users [Anderson et al., 2020]. However, the new engagement may be inconsistent with the norms of the community which could have a negative externality on existing users [Gu et al., 2007].

In addition, our findings on the diversity of engagement have important policy and social implications. Previous work has found that personalized news feed algorithms and personalized recommendations have led to filter bubbles and decreased content consumption diversity [Allcott et al., 2020, Bakshy et al., 2015b, Claussen et al., 2019, Holtz et al., 2020, Levy, 2021]. Moreover, these studies have found a tradeoff in engagement quantity and diversity that led \cite{holtz2020engagement} to introduce the concept of the "Engagement-Diversity Connection." We demonstrate that augmenting personalized feed algorithms with non-personalized feeds of news related content can increase both engagement and engagement diversity. This is true both for engagement among different categories of news as well as for the diversity of engagement among publishers of different political slant. From a managerial perspective, more diverse engagement has been shown to positively impact user retention [Anderson et al., 2020]. From a policy perspective, users consuming and engaging with a more diverse set of political viewpoints has important positive implications for civic life [Sunstein, 2003]. This suggests that non-personalized feeds can be an important tool in mitigating algorithmic filter bubbles and should be studied further to assess the robustness of these findings.

Chapter 3

Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology[1]

“We should stop training radiologists now. Its just completely obvious that within five years, deep learning is going to do better than radiologists."

– Geoffrey Hinton (in 2016)

3.1 Introduction

Artificial intelligence (AI) is a general-purpose technology with transformative potential similar to that of the steam engine and electricity [Acemoglu and Johnson, 2023, Agrawal et al., 2018, Brynjolfsson and Mitchell, 2017, Brynjolfsson et al., 2017, Frank et al., 2019, Goldfarb et al., 2023]. But, in contrast to the innovations of the industrial revolutions, AI can perform tasks that require complex reasoning (Webb, 2019; Felten et al., 2019; Brynjolfsson and Mitchell, 2017). Indeed, a growing literature shows that AI can outperform humans in a host of predictive tasks, including those typically performed by experts (Liu et al. 2019; Lai et al. 2021; Mullainathan and Obermeyer 2019; Kleinberg et al. 2017; Agrawal et al. 2018).[2]

Radiology is as an iconic example of this development. Yet, many disagree with Hinton's proclamation that AI will replace radiologists.[3] These skeptics argue that instead of human

[1] This chapter is written in collaboration with Nikhil Agarwal, Tobias Salz, and Pranav Rajpurkar [Agarwal et al., 2023].

[2] We will use the term AI to refer to a neural net-based image classifier. The term artificial intelligence is typically reserved for a system of different prediction tasks to mimic a more complex set of behaviors, whereas machine learning is concerned with one specific prediction task. For a detailed discussion of this distinction see [Taddy, 2018], among others.

[3] A more nuanced but qualitatively similar prediction that machine learning tools will displace radiologists

radiologists being replaced by AI, it is optimal for them to use AI assistance [Agrawal et al., 2019, Langlotz, 2019]. In addition to considerable legal and regulatory challenges that stand in the way of full automation, combining human expertise with AI input has potential gains that cannot be realized by exclusively relying on one or the other. For example, radiologists may correct mistaken AI predictions or may have access to information about the clinical context on which the AI is not yet trained. Current regulatory practice by the FDA is consistent with these arguments: approved AI tools for clinical decision-making typically play a supporting role rather than operating autonomously [see Harvey and Gowda, 2020, Norden and Shah, 2022, for example]. Similar arguments can be made in many other settings where AI approaches or exceeds the abilities of human experts.

This paper investigates the optimal form of collaboration between humans and AI. That is, should AI predictions that surpass human performance be used to automate decisions or to assist humans? The answer to this question depends on our three broad questions. First, do humans hold valuable information not included in AI predictions? If yes, then one would like to harness this information by using AI to augment humans instead of fully automating decisions. However, a substantial literature in economics suggests that humans may err when making probabilistic judgments by deviating from the benchmark model of Bayesian updating with correct beliefs [see Benjamin et al., 2019, for a review]. In the presence of such mistakes, it may not be optimal to always give the human access to the AI's information. This brings us to the next two questions: How do humans combine AI predictions with their own information? And how do potential mistakes shape the optimal form of human-AI collaboration?

We design and run an experiment with professional radiologists and develop an empirical methodology that aims to answer these questions.[4] Our experimental design compares human and AI performance, quantifies the predictive value of the information that humans hold but AI tools do not (henceforth termed contextual information), and tests whether AI assistance improves human performance. We then develop a method to estimate a model of (potentially imperfect) belief updating, analyze what this model implies about the optimal form of collaboration between AI and humans, and apply it to our experimental data.

The experiment includes 227 professional radiologists recruited through teleradiology companies to diagnose retrospective patient cases. Radiology offers an environment that is both naturalistic and allows us control similar to that in a laboratory experiment. As in our experiment, radiologists often work remotely, and our interface resembles the one they typically use. Our treatments vary the information set radiologists have access to when

is conveyed in Obermeyer and Emanuel [2016].

[4] We will use the terms 'humans', 'radiologists', and 'participants' interchangeably.

making decisions, in a two-by-two factorial design. In the minimal information environment, we provide only the chest X-ray image to which we add either AI predictions, contextual information, or both. Using an algorithm trained on about 250,000 X-rays with corresponding disease labels, the AI information treatment provides probabilities that a patient case is positive for a potential chest pathology [Irvin et al., 2019]. This algorithm was shown to perform comparably to board-certified radiologists. The contextual information treatment provides clinical history information that radiologists typically have available but, for data privacy reasons, is difficult to obtain to train the AI. This information includes the treating doctors' indications, the patient's vitals, and the patient's labs.

We will evaluate the quality of assessments by both AI and our participants against a diagnostic standard for each patient case. We follow the machine learning literature [Sheng et al., 2008] and construct a diagnostic standard by aggregating the assessments of five board-certified radiologists practicing at a highly reputed hospital with at least ten years of experience and chest radiology as a sub-specialty.[5] We also assess the robustness of all our results by constructing a (leave-one-out) diagnostic standard using the assessments of our experimental participants and by varying the aggregation method. Although this standard may not perfectly capture a "ground truth," patients would likely benefit from a system that brings diagnoses closer to the aggregate opinion of several highly qualified and experienced experts.[6]

We use the experimental data to estimate the value of contextual information and AI assistance, unpack biases in how humans use AI assistance, and analyze the optimal delegation problem. First, we estimate the treatment effects of our informational interventions on radiologists' prediction quality and the probability of making a correct decision. Next, we analyze whether and how humans deviate from a Bayesian benchmark when incorporating AI predictions. For example, humans may suffer from automation bias, a tendency to place more weight on machine-provided predictions than on one's own information.[7] Additionally, humans may treat AI predictions as independent of their own information [Enke and Zimmermann, 2019]. We show what different types of deviations from Bayesian updating imply for the collaboration between humans and AI. Finally, we quantitatively evaluate the optimal human-AI collaboration in terms of diagnostic performance and costs of human time. We assume that the AI signal can always be obtained at zero marginal cost and implement

[5] Unfortunately, medical records are of limited value because definitive diagnostic tests do not exist for most thoracic pathologies and, even when they do exist, are selectively performed depending on a radiologist's recommendation.

[6] A similar motivation justifies the use of second opinions in medical care.

[7] This terminology is borrowed from the literature that dates to the proliferation of computerized automated support systems in aviation, research which raised concerns about human complacency or automation bias [see Alberdi et al., 2009, for an overview].

a classifier that decides, as a function of the AI prediction, to delegate a case to either a human, a human with access to the AI, or the AI alone.

There are two key empirical challenges that we address through a combination of experimental designs. First, due to the high cost of recruiting radiologists at market rates, an across-participant design is impractical to power, except for very large effect sizes. We address this issue by adopting a within-participant design, where participants are randomized to experience four informational environments in random order, avoiding repeated case encounters. Second, to estimate a model of belief updating, it is important to obtain radiologists' diagnoses with and without AI assistance. Our second experimental design therefore asks participants to assess each case in each of the four information environments, with at least a two-week pause between repetitions of a case to minimize memory and anchoring biases. To ensure that our results do not rely on this "wash-out" being successful, a third design obtains an assessment with AI assistance only after assessments without AI assistance have been obtained. However, this third treatment is subject to order effects. We find no evidence of order effects on diagnostic quality although there is evidence that familiarity with the interface increases the speed with which participants go through patient cases. Our treatment effect analysis uses data from all designs whereas our model estimate of belief updating only uses data in which a radiologist reads the same case both with and without AI assistance.

We find that AI assistance does not improve humans' diagnostic quality on average even though the AI predictions are more accurate than approximately 75% of the participants in our experiment. Moreover, the zero average effect cannot be explained by the participants ignoring these predictions – we observe that radiologists' reported probabilities move significantly towards AI predictions when AI assistance is provided. Instead, the zero effect of AI assistance is driven by heterogeneous treatment effects: diagnostic quality increases when the AI is confident (i.e. the predicted probability is close to zero or one) but decreases when the AI is uncertain. In parallel, AI assistance improves diagnostic quality for patient cases in which our participants are uncertain, but decreases quality for patient cases in which our participants are certain. In contrast, providing clinical history does improve diagnostic quality, a result that suggests humans have additional valuable information that has not yet been incorporated into AI predictions.

An upshot of the results is that information available only to radiologists is useful, but humans do not correctly combine their information with AI predictions. In fact, AI predictions reduce predictive preformance for a range of signals. This result cannot be rationalized if our participants are Bayesians with correct beliefs because the AI assistance provides weakly more information to the decision-maker.

Motivated by these findings, we analyze two types of deviations from the benchmark model with correct updating to link errors in probabilistic judgement and optimal deployment of AI assistance.[8] The first type of deviation occurs when agents do not put the correct relative weight on the AI information. We describe this deviation using the approach introduced in Grether [1980, 1992] [see Benjamin, 2019, for a review] to define biases in belief updating. We say that an agent exhibits automation bias if they over-weight the AI information relative to their own and automation neglect if they under-weight it. The second type of deviation occurs if agents utilize an incorrect joint distribution of their own information and AI information; an example of such a deviation is correlation neglect [Enke and Zimmermann, 2019]. Our theoretical analysis shows that if agents exhibit only automation neglect, then AI assistance unambiguously increases diagnostic quality. All other forms of biases we consider result in AI assistance reducing diagnostic quality for certain realizations of AI and own information.

We then develop a method and use the data from our experiment to estimate empirical analogs of the deviations described above and select the model that best describes the treatment effects we document. This exercise requires us to solve several challenges unique to a naturalistic setting. One of the hurdles in our setting is that we, unlike in a laboratory game, cannot control the distribution of AI predictions and human information in our experiment, which differs from prior empirical applications of Grether's model of which we are aware.

In the model that best describes the data, agents exhibit automation neglect and act as if their own information and AI predictions are independent (conditional on the truth), even though this is not the case. Although parsimonious, we find that this model replicates the empirical patterns observed in the data. An important implication of the model is that it is not optimal to always provide AI assistance.

Thus, we turn our attention to designing a human-AI collaborative system that can selectively use AI predictions. We start by estimating the trade-off between diagnostic quality and radiologist time when, as a function of AI predictions, the diagnosis of a case's pathology can either be delegated to a human with or without AI assistance or be fully automated. The data from our experiment allow us to compute both of these quantities for each mode of diagnosis.

The results from this exercise mirror our treatment effect analysis: because radiologists take more time with AI assistance and do not correctly incorporate the AI's information, the majority of cases are optimally decided either by the radiologist or the AI alone but not by the radiologist with access to AI. We also find that signficantly more cases would have

[8] We will remain agnostic about whether the deviations we consider are due to non-Bayesian updating or can be explained by Bayesian updating with an incorrect mental model of AI predictions.

been optimally diagnosed by a human with AI assistance if humans correctly combined AI predictions with their own information, thus pointing to the potential importance of learning or further training.

Related Literature

A growing body of literature in computer science has explored the predictive performance of humans versus machine learning algorithms, with radiology often serving as a key area of application [Rajpurkar et al., 2017, 2018]. The study of human-AI collaboration has also become an increasingly important facet of medical AI research [Reverberi et al., 2022, Tschandl et al., 2020]. For comprehensive overviews of these areas, see Hosny et al. [2018], Lai et al. [2021], Rajpurkar et al. [2022], Zhou et al. [2021]. Research on the effectiveness of human-AI collaboration is evolving, with notable studies in radiology including Fogliato et al. [2022], Kim et al. [2020], Park et al. [2019], Rajpurkar et al. [2020], Seah et al. [2021]. An active literature studies whether AI assistance benefits radiologists, and which radiologists benefit the most [Ahn et al., 2022, Gaube et al., 2023, Rajpurkar et al., 2020, Seah et al., 2021, Sim et al., 2020]. Another set of papers build delegation algorithms to predict the types of cases for which human performance exceeds machine performance [e.g. Bansal et al., 2021, Mozannar and Sontag, 2020, Raghu et al., 2019]. In contrast to prior studies, we recruit a large group of high-skilled experts under contracts that allow us to incentivize our participants. A key conceptual difference is that, unlike previous studies which are mainly concentrated on performance, our work emphasizes behavioral biases, how they can be measured in a naturalistic setting, and their impact on human-AI interaction and optimal AI deployment.

A rapidly growing literature in economics also compares human and AI performance. Within economics, these studies tend to rely on observational approaches, with examples addressing issues in medicine [Mullainathan and Obermeyer, 2019, Ribers and Ullrich, 2022] and bail decisions [Angelova et al., 2022, Kleinberg et al., 2015], amongst others. However, analyses based on observational data face critical identification challenges, such as the selective labels problem [see Kleinberg et al., 2017, Mullainathan and Obermeyer, 2019, Rambachan, 2021]). A limited set of studies use quasi-experimental approaches (e.g., Angelova et al., 2022, Stevenson and Doleac, 2019) or randomized controlled trials (e.g., Bundorf et al. [2020], Grimon et al. [2022], Imai et al. [2020], Noy and Zhang [2023]) to investigate human use of AI tools, typically focusing on overall performance or variability in participant response. We add to this literature by developing an experimental approach that manipulates the information environment that calculates and compares behavior with a Bayesian benchmark to document systematic biases and demonstrate that these biases lead to a non-trivial delegation problem.[9]

While several studies in behavioral economics have documented errors in probabilistic judgment and belief formation, they do not consider the consequences for AI deployment [c.f. Benjamin et al., 2019, Conlon et al., 2022, Enke and Zimmermann, 2019, Tversky and Kahneman, 1974, for example]. Our definitions of automation bias builds on the framework in Grether [1980]. We contribute to this literature in two ways. First, we develop an approach to estimate the parameters of the model in Grether [1992] in an environment where the joint distribution of the signals cannot be controlled (or partialled out) by the researcher.[10] This methodological advance is necessary because we cannot modify the signal within medical images. Second, we link the design of AI information provision to the (biased) updating rule that humans use. This link shows that utilizing AI information by humans is an important and practical application of the ideas in this literature.

Finally, our work also adds to the literature on decision-making, particularly in the health care context [e.g. Abaluck et al., 2016, Chan et al., 2022, Chandra and Staiger, 2020, Currie and MacLeod, 2017, Gruber et al., 2021]. Such efforts use observational data on medical decisions to understand predictions and payoffs, objectives that are achievable under less stringent functional form restrictions in our experimental approach. An important distinguishing feature is that none of these papers consider the effects of AI predictions.

Overview

The rest of the paper is organized as follows. Section 3.2 introduces our model of a decision-maker in a diagnostic setting. Section 3.3 describes the necessary details of the setting and our experimental design. Section 3.4 discusses the treatment effects. Section 3.5 estimates a descriptive model of deviations from Bayesian updating. Section 3.6 shows the gains achievable under the optimal collaboration between radiologists and AI.

3.2 Conceptual Model

Our study focuses on classification problems and prediction algorithms intended for these tasks. These algorithms are designed to predict the appropriate classification for a given case and may assist a human decision-maker. This decision-maker, indexed by h, must

[9] Our finding that radiologists exhibit automation neglect is related to those in Dietvorst et al. [2015], which shows that humans are averse to following algorithmic recommendations as compared to human recommendations. This aversion can be reduced if humans are allowed to modify the algorithm's recommendation [Dietvorst et al., 2018].

[10] Most applications that we are aware of rely on one of two experimental approaches. In the first approach, the researcher can partial out either the prior information or the likelihood ratio of the signal provided, for example in the classic bookbag-and-poker-chip experiments (see Benjamin et al. 2019, Benjamin 2019, for reviews). In the second approach, the researcher directly provides signals from a known joint distribution (see Conlon et al. [2022]).

take a binary action $a_{ih} \in \{0,1\}$ on case i based on a prediction of a binary class $\omega_i \in \{0,1\}$. The realized payoff $u_h(a_{ih}, \omega_i)$ from an action depends both on the correct class and the action. The human does not know ω_i but observes a subset of two signals that are potentially informative about the state depending on the information environment. The first signal is generated by a prediction algorithm (AI), with realizations $s_i^A \in S^A$. The second signal is directly obtained by the human, with a realization $s_{ih}^H \in S^H$. These signals are of arbitrary dimension. The joint distributions of the signals conditional on the state is given by $\pi_h(\cdot|\omega) \in \Delta\left(S^A, S^H\right)$, with prior probabilities over the class $\pi(\omega)$. We do not place any restrictions on $\pi_h(\cdot|\omega)$ – the signals need not be independent conditional on the state of the world, the signal distribution may depend on the human to capture skill heterogeneity [Chan et al., 2022], and one of the signals could be more informative than the other [Blackwell, 1953].

Assume that the human's objective is to correctly classify each case. It is without loss of generality to normalize the payoff from taking the action that matches the correct class to zero. Let $c_{FP,h}$ be the disutility of human h if they set $a = 1$ when $\omega = 0$ (false positive) and $c_{FN,h}$ be the disutility if they set $a = 0$ when $\omega = 1$ (false negative). The payoff of the human is therefore

$$u_h(a,\omega) = -1 \cdot \{a = 1, \omega = 0\} \cdot c_{FP,h} - 1 \cdot \{a = 0, \omega = 1\} \cdot c_{FN,h}. \tag{3.1}$$

We allow the human's posterior belief given the observed signals to deviate from those implied by the true probability law $\pi_h(\cdot|\omega)$. Specifically, let $s_{ih} \subset \left\{s_i^A, s_{ih}^H\right\}$ be the subset of signal realizations observed by the human h and $p_h(\omega\,|s_{ih}) \in [0,1]$ be the human's belief when they observe s_{ih}. Suppressing the dependence of signals on the pair (i,h), the human's action given the signal s is

$$a_h^*(s;p_h) = 1 \cdot \left\{\frac{p_h(\omega = 1\,|s)}{p_h(\omega = 0\,|s)} > c_{rel,h} \equiv \frac{c_{FP,h}}{c_{FN,h}}\right\}. \tag{3.2}$$

The expected payoff from following $a^*(s)$ is

$$V_h(s;p_h) = E\left[u_h\left(a_h^*(s;p_h), \omega\right)|s\right] = \sum_\omega u\left(a_h^*(s;p_h), \omega\right)\pi_h(\omega\,|s),$$

where decisions are based on the human's belief p_h, but are evaluated according to the true law π_h. Because we allow for p_h to differ from π_h, the action $a_h^*(s;p_h)$ can deviate from the optimal action $a_h^*(s;\pi_h)$ given the signal $s = \left(s^A, s^H\right)$. Except in knife-edge cases, $V_h(s;p_h)$ is lower than $V(s;\pi_h)$ whenever $a^*(s;p_h) \neq a^*(s;\pi_h)$.

The discussion above shows that the effect of AI assistance on decision quality depends on whether humans' beliefs with AI assistance deviate from the benchmark given by Bayesian updating with correct beliefs (about the joint distribution of the signals and the correct class). Bayes' rule implies that, given the signals $\left(s_i^A, s_{ih}^H\right)$, the decision-relevant log-odds is given by

$$\log \frac{\pi_h\left(\omega_i = 1 \left| s_i^A, s_{ih}^H\right.\right)}{\pi_h\left(\omega_i = 0 \left| s_i^A, s_{ih}^H\right.\right)} = \log \frac{\pi_h\left(s_i^A \left| \omega_i = 1, s_{ih}^H\right.\right)}{\pi_h\left(s_i^A \left| \omega_i = 0, s_{ih}^H\right.\right)} + \log \frac{\pi_h\left(\omega_i = 1 \left| s_{ih}^H\right.\right)}{\pi_h\left(\omega_i = 0 \left| s_{ih}^H\right.\right)}, \tag{3.3}$$

where the second term on the right-hand side is the posterior log-odds ratio for the two states $\omega_i = 1$ to $\omega_i = 0$ given that the human's signal is s_{ih}^H. Thus, one goal of our exercise is to estimate the left hand side and compare it with the analogous quantity for $p_h(\cdot)$.

Estimating the benchmark odds ratio above is empircally is challenging even if $p_h(\cdot)$ can be elicited because of two conceptually important reasons. The first challenge is that signals s_{ih}^H and (correct) beliefs $\pi_h(\cdot)$ differ across patient cases i and across humans h because of patient and radiologist skill heterogeneity respectively.

The second challenge arises because constructing the terms on the right hand side of equation (3.3) requires a conditioning on s_{ih}^H. In fact, even though the first term on the right hand side is an update due to the AI signal, accounting for potential correlation with s_{ih}^H requires controlling for it. Unlike in some laboratory settings, we do not directly observe humans' signals. Our econometric approach, which is discussed in section 3.5, will construct controls from reported beliefs without AI assistance.

The ideal dataset for addressing these challenging would elicit beliefs given a human h and a case i both with and without the AI signal. However, empirically implementing this strategy requires us to eliminate concerns about anchoring and order effects when eliciting beliefs about the same case twice. We thus turn our attention to the experimental design that is aimed at solving these issues.

3.3 Setting and Experiment

Our experiment elicits the probability of a pathology's presence $p_h\left(\omega_i = 1 \left| s_{ih}\right.\right)$ and a recommended treatment/follow-up decision a_{ih} under varying information treatments. There are four information treatments in the experiment. In the minimal information environment participants only observe the chest X-ray, to which we add AI assistance, contextual information, or both. Next, we describe the experimental context and interface before presenting the design of our experiments.

3.3.1 Experimental Context

Radiology

Radiologists diagnose the presence of a given pathology at the request of a treating physician. The information available to a radiologist consists of diagnostic images (e.g. chest X-rays), any relevant medical history (e.g. laboratory results), and clinical indication notes of the treating physician. The treating physician's notes are of varying detail levels – they may provide no clinical information or guidance, request the analysis of a specific pathology, or only list the patient's primary symptom (see appendix C.2.2 for examples). Radiologists are expected to report all pathological findings irrespective of the pathology suspected by the treating physicians.

Because image-based classification is a core task performed by radiologists AI tools have made significant inroads in the field. Recent advances in deep learning methods for image recognition have yielded algorithms that can match or surpass the performance of human radiologists [Langlotz, 2019, Obermeyer and Emanuel, 2016]. As of 2020, 55 companies offered a total of 119 algorithmic products of which 46 have FDA approval [Tadavarthi et al., 2020]. Most products related to clinical decision-making are marketed as support tools as opposed to autonomous tools, partly due to regulatory and liability issues [Harvey and Gowda, 2020].

CheXpert

We provide AI assistance using predictions from the CheXpert model, which is a deep learning prediction algorithm for chest X-rays [Irvin et al., 2019]. This model is trained on a dataset of 224,316 chest radiographs of 65,240 patients labeled for the presence of fourteen common chest radiographic pathologies. The algorithm does not use any other patient information, such as the clinical history or vitals.[11] Nonetheless, a prior version of this algorithm was shown to match or surpass the performance of board-certified radiologists from Stanford Hospital on five pathologies [Patel et al., 2019]. These results are also presented to our participants when introducing the AI tool. Section 3.4 confirms that the algorithm outperforms a majority of radiologists in our experiment. We relegate additional details about the algorithm to appendix C.2.3. The algorithm assistance to our participants will be in the form of a vector of probabilities for the presence of each CheXpert pathology.[12]

[11] While large datasets of images are increasingly available (e.g. Kramer et al. 2011, Johnson et al. 2016, Irvin et al. 2019) it is significantly more difficult to construct such datasets for other patient information due to the compulsory manual review of textual data for HIPPA compliance.

[12] Some algorithms attempt to make their predictions explainable to a human by highlighting the parts of the image that drive a specific prediction. However, prior studies show that providing such localization in

3.3.2 Experimental Designs

Our experiment varies the information available to diagnose patient cases—participants may or may not receive AI assistance and may or may not have access to the clinical history. The X-ray is shown under all information conditions. We expose our participants to all four possible information conditions: X-ray only, henceforth *XO*; clinical history without AI, henceforth *CH*; AI without clinical history, henceforth *AI*; and both clinical history and AI, henceforth *AI+CH*.

There are two objectives of our experiment. The first is to compute the treatment effects of AI and CH on diagnostic quality and radiologist time. The second is to analyze how radiologists update when receiving the AI signal and compare it to a Bayesian benchmark.

Both these objectives are complicated by the likely heterogeneity in radiologist skills. For estimating treatment effects, radiologist heterogeneity implies that a design that randomizes treatments only across radiologists will require a large participant pool except for extremely large effect sizes. Our participants are highly paid, making this approach expensive. And, as explained in section 3.2, across-radiologist variation in information treatments is not tailored for the second objective. We would ideally know how a given radiologist changes her assessment for the same case under a different information condition.

Our approach to address these challenges is to use a combination of three different experimental designs, each with certain advantages and disadvantages. Appendix C.2.1 illustrates the three design variations.

Design 1 (Figure C.1)

In the first design, participants are assigned to a random sequence of the four information treatments. Each information condition is assigned fifteen cases at random without repetition. Participants read all 15 cases in one information environment before moving to the next one.

This design builds in both across- and within-participant variation in information treatments. The within-participant variation has greater power because it controls for participant heterogeneity at the potential cost of order effects. The concern of order effects is both testable and mitigated by the randomization of treatment sequence across subjects.

This first design is well-suited to estimate treatment effects of our information environments. However, as mentioned earlier, it is not ideal for estimating an empirical analog to equation (3.5) because no case is encountered twice.

addition to the numeric output does not improve the accuracy of radiologists [Gaube et al., 2022]. Moreover, a quantitative output allows us to compute a Bayesian benchmark to the radiologist's prediction, which is otherwise difficult.

Design 2 (Figure C.2)

Radiologists diagnose each patient case in each of the four information environments in the second design. For the moment, set aside concerns arising from the feature that the same radiologist encounters the same case multiple times. This design will allow us to estimate an empirical analog to equation (3.5). It also has the added benefit of controlling for both case-radiologist heterogeneity because, unlike in the previous design, we can conduct within-case-radiologist comparisons across treatments.

Because radiologists repeatedly encounter cases, we need to address the potential for order effects due to memory. For example, radiologists might anchor on their previous assessment using AI predictions or contextual information and might remember this information the next time the same case is encountered. We, therefore, limit radiologists' ability to remember either their diagnosis or previously provided information by using a "washout" interval between two encounters of the same case.[13] Specifically, radiologists complete the experiment in four sessions that are separated by at least two weeks. Each session is similar to the first design: radiologists diagnose fifteen cases in each of the four information environments with no case repeated within a session. Across sessions, the information environment under which a given case is diagnosed is permuted. Thus, by the end of the fourth session, each of the sixty cases is diagnosed exactly once in each information environment. Our results are consistent with the washout being effective – radiologists' predictions do not move towards the AI prediction if it was provided in a prior session but do if it is provided in the current session (see figure C.37).

Design 3 (Figure C.3)

In the third design, we address residual concerns about the order effects of radiologists diagnosing cases with AI before those without AI–whether due to anchoring, memory, or experimenter demand–by having participants diagnose fifty cases, first without and then with AI assistance. Within each block, clinical history is randomly provided in either the first or second half of images.

This design also allows us to conduct within case-radiologist comparisons. The potential disadvantage of this design is that we cannot distinguish order effects from the effect of providing AI. This issue is unavoidable given the guiding principle that participants receive weakly more information about a case during a repeat encounter. However, we can test for and do rule out order effects on accuracy based on the first two designs.

[13] This principle has been used in computer science [Conant et al., 2019, Pacilè et al., 2020, Seah et al., 2021].

Participant Recruitment

Participants for the first and third designs, which constitute the majority, were recruited through teleradiology companies. Most healthcare providers in the US rely on these companies' services, even those that have on-call radiologists [Rosenkrantz et al., 2019]. We work with teleradiology companies that serve US hospitals and offer the services of both US-based and non-US-based radiologists. Our contract specifies a piece-rate, and the companies, in turn, compensate the participants with a piece-rate.[14] In addition, we provided monetary incentives for accuracy to a subset of radiologists, as described in the next section.

The second design required us to work with a partner who could guarantee subjects' participation over several months. We collaborated with VinMac healthcare system in Vietnam to recruit their staff radiologists to ensure continued participation. VinMac is in the process of developing its own in-house AI capabilities and was willing to assist with our experiment in exchange for recognition in a publication of the resulting dataset. The VinMac radiologists did not receive receive any payments to participate in the experiment but we find that their perfomance is very close to the performance of the tele-radiologists.

In total, 227 radiologists participated in our experiment. Approximately 14% of our participants are US-based, 15% have a degree from a US institution, 44% are affiliated with a large clinic, and 63% with an academic institution. As demonstrated in appendix C.3.4, the quality of the assessments made by the radiologists in our study is comparable to that of the staff radiologists from Stanford University Hospital, who originally diagnosed the patient case.

Incentives

We cross-randomize incentives for accuracy in the first and third designs but not the second because of the specific ways in which our partner's radiologists are employed. Payments were determined following the binarized scoring rule in Hossain and Okui [2013], where truth is determined as described in section 3.3.3 below. This incentive scheme uses a loss function of the mean squared prediction error, averaging over patient cases and pathologies, and the respondents earn a fixed bonus of $120 if a random draw is less than the loss function. This bonus is more than 20% of the base payment to teleradiology firms. We explain to the participants that expected payments are maximized if they provide their best estimates using a non-mathematical description of the payment rule. We specify the distribution so that 30% of pilot participants would earn the bonus, cross-randomized with the other two treatment arms.

[14] The piece-rate we pay the teleradiology companies range from $7.50 to $13.00.

3.3.3 Implementation and Data Collection

Patient Cases and Diagnostic Standard

The experiment uses 324 historical patient cases with potential thoracic pathologies from Stanford University's healthcare system. For each case, we have access to the chest X-ray and the clinical history in the form of the primary provider's written notes, the patient vitals, and demographics.[15] The use of retrospective cases allows us to avoid ethical and other issues that would arise when experimenting in high-stakes settings.

Our analysis requires constructing the correct class ω_i for each patient case and pathology. We construct ω_i by aggregating the assessment of a group of expert radiologists, an approach common in computer science [Mccluskey et al., 2021, Sheng et al., 2008]. We asked five board certified radiologists from Mount Sinai with chest specialty to read each of the 324 cases using the interface described above with the available X-ray and clinical history. For each case-pathology i and radiologist h, we obtain $\pi_h\left(\omega_i = 1 \middle| s^H_{i,h}\right)$. We classify $\omega_i = 1$ if $\sum_h \pi_h\left(\omega_i = 1 \middle| s^H_{i,h}\right)/5 > 0.5$. We interpret ω_i as the diagnostic standard for a case-pathology given all available information at the time of diagnosis.

The diagnostic standard may differ from the "ground truth" presence of a pathology. However, obtaining such "ground truth" for an unselected sample of patient cases is infeasible in most diagnostic settings. Additional information in medical records are often inconclusive because definitive tests do not always exist, and follow up patient care and outcomes are selected based on the assessed presence of a pathology.[16] This issue is referred to as the selective labels problem [e.g. Mullainathan and Obermeyer, 2019]. Recent literature has suggested instrumental variables approaches for solving this selective labels problem, but this work targets population quantities and not a "ground truth" on each case [e.g. Chan et al., 2022, Mullainathan and Obermeyer, 2019].

In comparison, the diagnostic standard immediately addresses the selective labels problem because the availability of assessments is not selected on the likelihood of a pathology being present. Results in Wallsten and Diederich [2001] suggest that, under weak conditions that allow for measurement error in the reports and correlations across reports, the aggregate opinion of several experts is highly diagnostic as long as the experts are median unbiased.[17]

[15] All cases are first encounters with no prior X-ray as a comparison. We started with 500 cases that fit these primary criteria. We omitted pediatric cases from this set. Finally, a radiologist reviewed the cases to remove instances with poor image quality. The clinical history was manually reviewed to remove patient-identifiable information and cleared for public release.

[16] Many pathologies do not have commonly used non-imaging-based diagnostic tools. For instance, the presence of cardiomegaly – an enlarged heart – can only be determined using imaging tools, thoracic surgery or an autopsy.

[17] Previous work cautions that physician opinions could reflect systematic underlying physican racial and

To assess robustness of our results, we consider several alternative constructions of the diagnostic standard and analyze a subsample of cases for which the standard is not ambigious. These variations are discussed after the baseline results.

Experimental Interface and Data Collected

We developed the experimental interface to present the patient cases and to collect radiologists' predictions and decisions in collaboration with board certified radiologists at Stanford University Hospital and Mt. Sinai Hospital. In contrast to free-text reports, we designed it to generate structured and quantitative data that resemble a typical radiological report. We briefly describe this interface and provide images and further details in appendix C.2.

On the landing page of each case, a high-resolution image of a patient's X-ray is presented to the radiologist, with the functionality to zoom and adjust brightness and contrast. When the experiment calls to show the clinical history, the interface presents clinical notes, vitals, and laboratory results available at the time the X-ray was originally ordered. If the experiment provides AI assistance, participants are shown AI predictions.

The probability that a pathology is present given the available information is elicited using a continuous slider. We visually subdivide possible responses into five intervals with standard language labels used in written radiological reports to aid the participants.[18] We also collect a binary "treatment/follow-up" recommendation for each pathology that is not definitively ruled out.[19] We will interpret this input as $a_h^*(s)$. In a real clinical setting, a recommendation to follow-up could trigger the treating physician to prescribe additional medical tests or interventions with potential costs and benefits. Thus, an optimal recommendation trades off the cost of false positives and false negatives when recommending an action as in section 3.2. The probabilistic assessments with the follow-up decision will allow us to estimate radiologists' relative cost of false positives and false negatives.

We elicit responses for pathologies in a hierarchical structure designed by our collaborating radiologists.[20] There are eight mutually exclusive top-level pathologies. For instance, "airspace opacity" is distinct from a "cardiomediastinal abnormality." Each of these top-level

other biases (see Mullainathan and Obermeyer, 2017, for example).

[18] The specific labels are *"Not present","Very Likely", "Unlikely", "Possible", "Likely", and "Highly Likely".* Several radiological publications have suggested such standardized language for radiological reports. See for instance Panicek and Hricak [2016].

[19] The binary treatment/follow-up decision is only asked for pathologies where a follow-up is clinically relevant. This includes all pathologies with AI assistance.

[20] The hierarchical structure reduced the data entry burden on our participants, and we piloted the interface with several radiologists specializing in the interpretation of chest X-rays. The groups all correspond to a standard class of pathologies and prior clinical research on AI in chest X-Ray image classification has used similar hierarchies [see Seah et al., 2021, for example].

pathologies has children that are more specific, which may be further subdivided in some cases. In addition, we elicited an overall assessment of whether the radiologists considers the case normal or not. In the main text we focus on analyzing the two top-level pathologies with AI predictions and drop further subdivisions from the analysis. Our results are robust to including the lower-level pathologies in the analysis as we show in appendix C.3.5.

In addition to $p_h(\omega = 1 \,|s)$ and $a_h^*(s)$, we record active time, response times, and any clickstream data that results from the interaction with the interface.[21] The participants are not explicitly informed about this monitoring, and there are no explicit time limits. Our experiment runs remotely, and participants connect to a server, which hosts the interface and records responses.[22]

Participant Training

We train the participants using a combination of written instructions and a video. The materials provide an overview of the experimental tasks, the interface, and information about the AI assistance tool. The firms and the participants know that the research study involves retrospective patient cases. To train participants on the AI tool, we provide them with materials that explain the development of the algorithm, present metrics of its performance on various diseases, and summarize the algorithm's performance relative to radiologists based on prior research. In addition, we show the participants fifty example cases that show the X-ray and clinical history next to the AI output. The participants are informed that the algorithm only uses the chest X-ray to form predictions, and this knowledge is later tested in a comprehension question. After the instructions, participants answer eight comprehension questions, which they must answer correctly before proceeding to the experiment. We also include an endline survey. We do not directly interact with the subjects except to field questions about the experiment or provide tech support. The complete set of instructions is provided in appendix C.2.2.

[21] Active time is calculated based on the clickstream data to approximate the time spent actively working on the study. We exclude instances where a participant pauses the study which would substantially increase the noise in the time measures.

[22] This interface is browser-based and built using the o-tree framework Chen et al. [2016]. Since we are not directly communicating with our participants we also deploy a device fingerprinting service from fingerprint.com to ensure that there are no repeat participants.

3.4 Estimated Treatment Effects

3.4.1 Overall Performance of AI and Radiologists

This section focuses on measures of performance (deviation from diagnostic standard, incorrect decision), deviation from AI prediction, and measures of effort. Table 3.1 summarizes the data on these measures and sample sizes from our experiment. The main text focuses on the two top-level pathologies with AI predictions (Cardiomediastinal Abnormality and Airspace Opacity) but our results are qualitatively robust to the inclusion of all pathologies with AI predictions (see appendix C.3.5).[23] A unit of observation is a radiologist decision (or prediction) for a given patient and pathology.

Radiologists give the correct follow-up/treatment recommendation in 70% of case pathologies. On average, they spend ~2.8 minutes per case with large variability across cases. All summary statistics are very similar across the three expterimental designs. For instance, the average deviation from the diagnostic standard, which is defined as $Y_{iht} = |p_h\left(\omega_i = 1|\, s_{iht}\right) - \omega_i|$, for the three designs ranges from 0.212 to 0.232, and average active time ranges from 2.58 to 2.88 minutes. Other measures, such as the share of correct decisions $(a_{ih} = \omega_i)$ and the deviation from AI assessments $\left(Y_{iht} = \left|p_h\left(\omega_i = 1|\, s_{iht}\right) - \pi\left(\omega_i = 1|\, s_i^A\right)\right|\right)$ are also similar across designs.

Before discussing the treatment effects, we compare the performance of the AI to the distribution of baseline performance of participating radiologists using two different measures. The first measure (AUROC) is derived from the receiver operating characteristic (ROC) curve, which measures the trade-off between the false positive and the true positive rate of a classifier. It is an ordinal measure whose value ranges from 0.5 for a classifier that guesses randomly to 1 representing perfect classification. The second measure is the root mean squared error (RMSE), which is cardinal and a lower value indicates higher performance. To compute these, we pool the data for top-level pathologies with AI for each radiologist's reports and for the AI's prediction (see appendix C.3.2 for pathology-specific comparisons).

The results are shown in figure 3.1 and indicate significant heterogeneity in performance across radiologists as well as the scope for AI assistance to improve radiologist performance. The heterogeneity across radiologists aligns with findings from observational data [e.g. Chan et al., 2022]. According to the AUROC, the AI is more predictive than 78% of radiologists and according to the RMSE more predictive than 90% of radiologists. Thus, there is ample room for AI assistance to improve the performance of radiologists. In fact, a majority of

[23] These pathology groups and, unless otherwise noted, the subsequent analyses were pre-registered (see SSR Registration 9620).

Table 3.1: Summary statistics

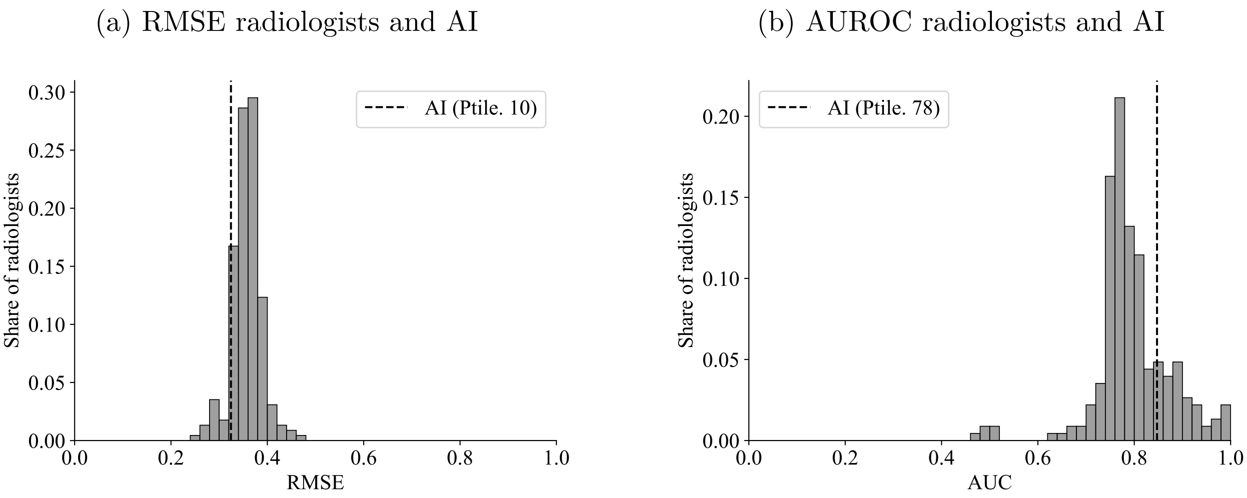
	All Designs		**Design 1**		**Design 2**		**Design 3**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reported Probability	0.229	0.289	0.211	0.285	0.245	0.278	0.240	0.322
Decision	0.311	0.463	0.268	0.443	0.400	0.490	0.231	0.421
Deviation from Diagnostic Standard	0.223	0.284	0.220	0.294	0.232	0.265	0.212	0.297
Deviation from AI	0.192	0.169	0.200	0.170	0.172	0.159	0.216	0.182
Correct Decision	0.704	0.456	0.745	0.436	0.620	0.485	0.785	0.411
Active Time	167.4	156.0	172.8	178.2	165.6	115.8	154.8	168.0
Observations	41,920		19,080		15,840		7,000	
Radiologists	227		159		33		35	
Reads per Radiologist	92.3		60		240		100	

Note: Summary statistics of the experimental data. Decision and accuracy statistics are for the two top-level pathologies with AI predictions (Cardiomediastinal Abnormality and Airspace Opacity) Columns (1) and (2) present the mean and standard deviation for all designs while Columns (3) and (4) present the same statistics for design 1 only, Columns (5) and (6) for design 2 only, and Columns (7) and (8) for design 3 only. Decision is an indicator for whether treatment/follow-up is recommended. Correct decision is an indicator for whether the decision matches the diagnostic standard. Deviation from diagnostic standard is the absolute difference between the reported probability and the diagnostic standard. Deviation from AI is the absolute difference between the human's reported probability and the AI's reported probability. Active time is measured in seconds.

radiologists would do better on average by simply following the AI prediction.[24]

[24] These results also align with Irvin et al. [2019], which shows that the CheXpert model yields a better classifier than two out of three radiologists on five pathologies and all three on three pathologies. Our results may differ from that because we use a different pool of radiologists, a different sample of cases, and reads with contextual information (clinical history) to construct the diagnostic standard. The latter two differences raise the bar for the AI because they reflect differences in the data-generating process.

Figure 3.1: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

Note: Distributions of two different accuracy measures of radiologist assessments alongside the AI's accuracy. Both distributions are shrunk to the grand mean using empirical Bayes. The histograms include the two top-level pathologies with AI predictions (Cardiomediastinal Abnormality and Airspace Opacity) and each observation represents a measure calculated at the radiologist level. The dotted line is the measure of the AI algorithm for the corresponding distribution, where "ptile" is short for "percentile." Only the assessments where contextual history information is available for the radiologists but not the AI prediction are considered. AUROC is only defined for radiologists who encounter some positive cases, which includes the large majority of radiologists. Robustness by design and diagnostic standard definition can be found in appendix C.3.5 & C.3.5.

We also compare the performances of our participants and the radiologist who originally diagnosed each patient case in appendix C.3.4.[25] There is no discernible difference between the two groups, which is consistent with the hypothesis that radiologists participating in the study were of similar skill and exerted similar effort as the radiologists completing the original reads.

3.4.2 How do Radiologists Respond to AI and Contextual Information?

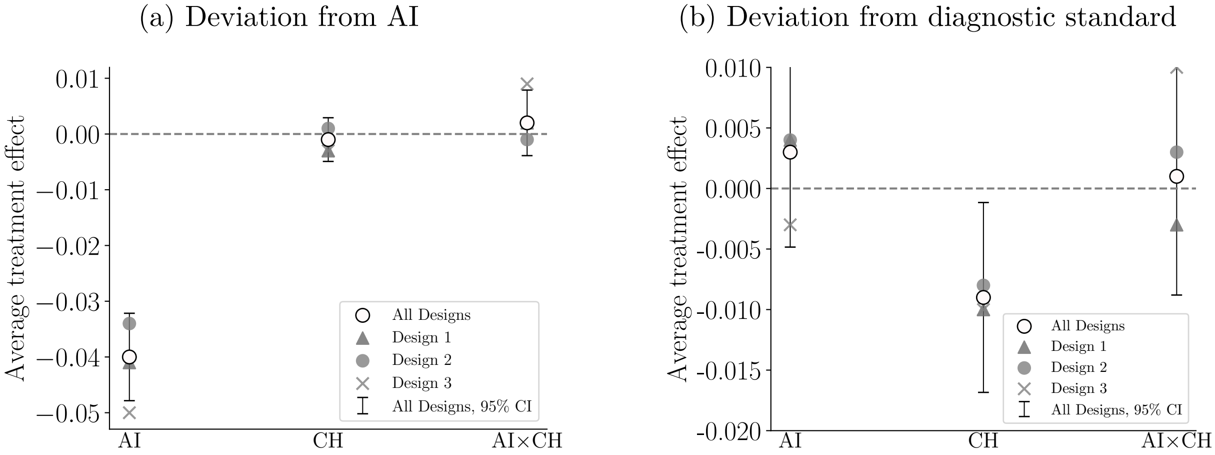
We now describe the effects of our information treatments estimated using the following specification:

$$Y_{iht} = \gamma_{g_i} + \gamma_{CH} \cdot d_{CH}(t) + \gamma_{AI} \cdot d_{AI}(t) + \gamma_{AI\times CH} \cdot d_{CH}(t) \cdot d_{AI}(t) + \varepsilon_{iht}, \tag{3.4}$$

where Y_{iht} is an outcome variable of interest for radiologist h diagnosing patient case-pathology i and treatment t, and γ_{g_i} are pathology fixed effects since there are multiple pathologies g_i for each case in this pooled analysis. Treatments t vary by whether or not clinical history is provided $d_{CH}(t) \in \{0, 1\}$ and whether or not AI information is provided

[25] We classified the original free text radiology reports associated with each case as positive, negative, or uncertain for each pathology using the CheXbert algorithm described in Smit et al. [2020]. To facilitate comparisons, we also discretized the probability assessments elicited during the experiment into positive and negative assessments. Then, we compared the accuracy of the original reads against the radiologists participating in the experiment.

Figure 3.2: Treatment effects of informational interventions



(a) Deviation from AI

(b) Deviation from diagnostic standard

Note: ATE of information treatments estimated using equation (3.4), on the deviation from AI (panel (a)) and the deviation from the diagnostic standard (panel (b)). Results are for the two top-level pathologies with AI predictions, airspace opacity and cardiomediastinal abnormality; separated by design and pooled across all designs. Standard errors are two-way clustered at the radiologist and patient-case level.

$d_{AI}(t) \in \{0,1\}$. We report two-way clustered standard errors at the radiologist and patient-case level. The estimates are robust to the inclusion of radiologist and patient-case fixed effects (appendix C.3.5). Cases are also balanced across treatments (see appendix C.3.1), which suggests that case randomization was successful. We will also compute conditional treatment effects given ranges of the AI signal s_i^A that are grouped based on $\pi\left(\omega_i = 1 | \, s_i^A\right)$.

Do Radiologists Utilize AI Predictions?

We begin by testing whether radiologists respond to the information that the AI provides. Panel (a) of figure 3.2 shows how the different information environments affect the disagreement of the radiologists' report with the AI's assessment measured using the deviation from AI. In calculating this deviation ($Y_{iht} = \left| p_h\left(\omega_i = 1 | \, s_{iht}\right) - \pi\left(\omega_i = 1 | \, s_i^A\right)\right|$), the term $p_h\left(\omega_i = 1 | \, s_{iht}\right)$ is the elicited probability whereas $\pi\left(\omega_i = 1 | \, s_i^A\right)$ is the AI's predicted probability that $\omega_i = 1$. When AI assistance is provided, then $s_{iht} = \left(s_{ih}^H, s_i^A\right)$, and otherwise $s_{iht} = s_{ih}^H$. The signal s^H also depends on whether contextual information is provided.

The results show that radiologists respond to AI assistance. Their predictions move significantly closer to the AI when receiving access to the AI prediction. To see this, observe that the control means for the deviation from the AI are approximately 0.21 for both when we pool designs and for design 1 only. Treatments where AI is provided reduce this baseline average deviation by 18%. We do not find a significant effect of clinical history on the deviation from the AI prediction nor do we find one from the interaction between AI and

CH.

Treatment Effects on Diagnostic Performance

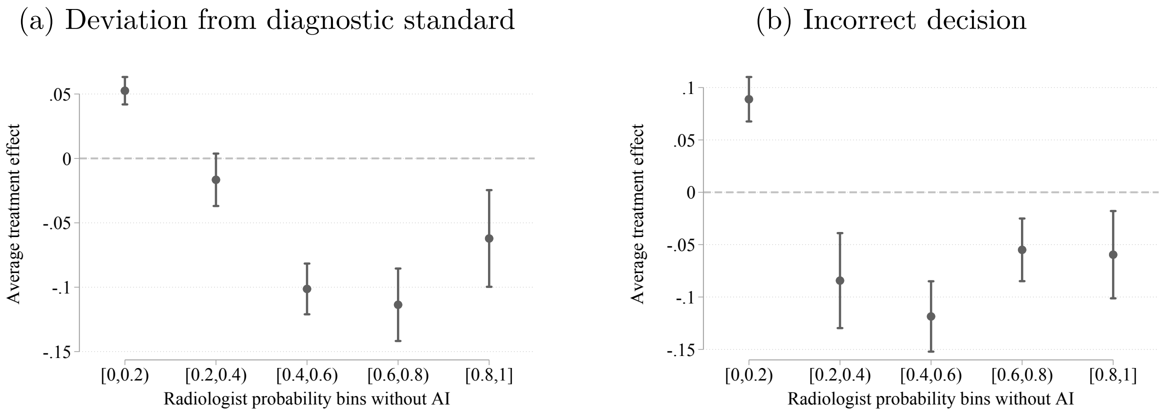
Next, we ask whether the information treatments affect radiologists' diagnostic performance measured using the deviation from the diagnostic standard. Recall that lower values imply better performance. Panel (b) of figure 3.2 also shows the average treatment effects on performance.

Access to contextual information improves performance on average. We find that access to clinical history reduces the deviation from the diagnostic standard by 4.0% ($p < 0.05$) of the control mean. This result suggests that one would like to utilize this information.

In contrast, AI assistance does not significantly improve average performance. The interaction between contextual information and AI assistance is also statistically indistinguishable from zero.

In light of the findings — that the AI is more accurate than most radiologists and that radiologists move their assessments toward the AI — it may seem puzzling that the AI information does not improve accuracy on average. This apparent contradiction occurs because the average treatment effects mask significant heterogeneity in treatment effects. Our within-participant designs — designs 2 and 3 — allow us to estimate conditional treatment effects given radiologists' predictions without AI assistance. Specifically, we partition cases based on the human's signal into five equally spaced bins of $p_h\left(\omega_i = 1|s_{ih}^H\right)$. Figure 3.3 shows the conditional treatment effects (pooled for design 2 and 3) of providing AI assistance on diagnostic performance. Panel (a) shows the deviation from the diagnostic standard and panel (b) shows the probability of incorrect decision. We find that providing AI assistance in cases when the radiologist is uncertain (i.e. the probability reported is not close to either zero or one) improves performance on both metrics, whereas AI assistance is harmful when the radiologist is close to certain that the pathology is not present for a given case.

Figure 3.3: Effect of AI by radiologist prediction without AI



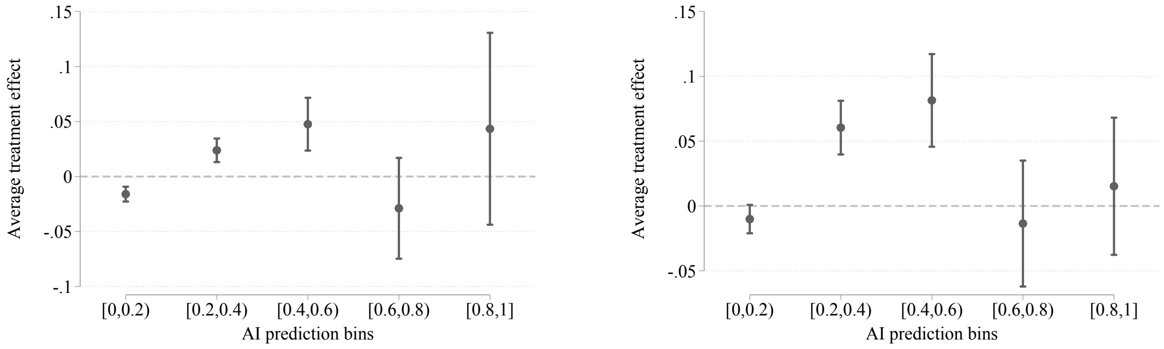
(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Panel (a) shows the conditional ATE of providing AI information on the deviation from diagnostic standard. Panel (b) shows analogous treatment effects on incorrect diagnosis. Standard errors are two-way clustered at the radiologist and patient-case level, with 95% confidence intervals depicted. Robustness to experimental design is in appendix C.3.5.

While AI assistance can help uncertain humans, we find that providing uncertain AI predictions reduces performance. As with the analysis of conditional treatment effects given human predictions, we estimate conditional treatment effects given AI predictions by partitioning cases into five bins based on $\pi\left(\omega_i = 1|s_i^A\right)$. Figure 3.4 presents the estimates, pooling data from all three experimental designs. When the AI provides a confident prediction (e.g. either close to zero or close to one) performance is significantly improved. We see that in the lowest bins of AI signals, the deviation from the diagnostic standard is reduced. In the second highest bin we also see a marked, though not statistically significant, improvement in performance. However, in the middle range of signals, where the confidence of the AI is low (meaning the AI signal is not close to either zero or one), radiologists' diagnostic performance and probability of making a correct decision is lower when AI information is provided.

Figure 3.4: Effect of AI by AI prediction



(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Panel (a) shows the conditional ATE of providing AI information on the deviation from diagnostic standard. Panel (b) shows analogous treatment effects on incorrect diagnosis. Standard errors are two-way clustered at the radiologist and patient-case level, with 95% confidence intervals depicted. Robustness to experimental design is in appendix C.3.5 and C.3.5.

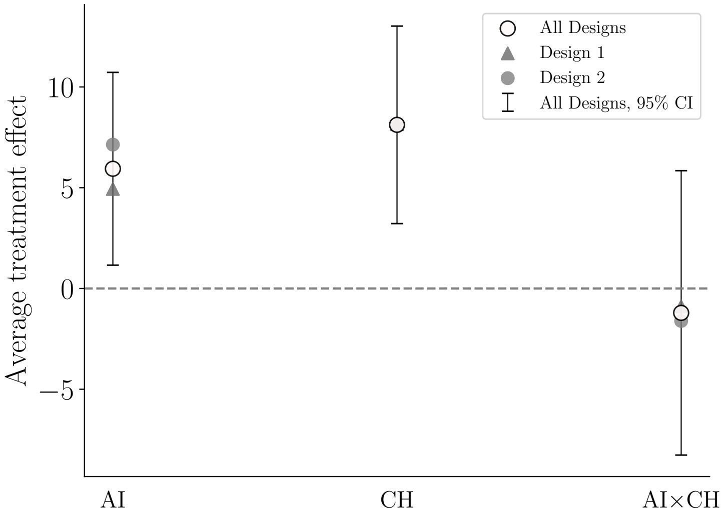
The results that AI assistance can decrease performance rejects a model in which radiologists are Bayesians with correct beliefs. Figures 3.3 and 3.4 showed that AI assistance reduces performance either if the radiologist was confident a pathology is not present or if the AI prediction is uncertain (i.e. the prediction is not close to either zero or one). Neither result can be rationalized in the benchmark model, suggesting that radiologists err when using AI predictions.

Treatment Effects on Time Per Case and Proxies of Effort

Finally, we turn our attention to the effects of AI assistance on time taken and the number of unique interactions (clicks) as proxies for effort. One hypothesis is that AI assistance could economize on costly human effort without sacrificing overall performance by enabling quicker assessments. Alternatively, it is possible that humans take more time because they are provided with more information to process. Which of these effects dominate determines the effect on labor costs when humans use AI assistance, and therefore the optimality of delegating cases versus a collaborative setup.

Our results indicate that radiologists are slower when provided with AI assistance. Figure 3.5 shows the treatment effects on time spent per case. These outcomes are measured at the case level. In the X-Ray Only treatment, radiologists spend about 2.6 minutes per case. Both AI and CH increase the time spend per case by a statistically significant amount of approximately 4%. The interaction term $\gamma_{AI \times CH}$ is not significant for either of the two outcome variables. These effects suggest that decisions where both radiologists and the AI

Figure 3.5: Effect of informational interventions on time



Note: ATE of the information treatments, estimated using equation (3.4), on radiologist effort measured in terms of active time (in seconds). Results exclude five patient-cases with an unaccounted time measure, and observations from design three because of potential order effects. Active time is winsorized to the 95th percentile. Standard errors are two-way clustered at the radiologist and patient-case level.

are involved come at a non-trivial increase in time spent per case. Treatment effects for clicks are displayed in appendix C.3.5. This result further undercuts the potential benefits in performance from including humans assisted with AI predictions "in the loop."

3.4.3 Robustness

Appendix C.3.5 shows that the results are qualitatively robust to a variety of alternative analyses. The treatment effect analysis in this section does not condition on the sequence in which subjects encounter information treatments. Reassuringly, they are statistically indistinguishable from those that use only an across participant comparison from the first treatment encountered in designs 1 and 2 (appendix C.3.5). Appendix also C.3.5 shows that our results are robust to including controls for order effects.

Alternative methods for constructing the diagnostic standard or focusing on cases with greater consensus also yield similar conclusions. The variations we consider include (i) using a leave-one-out diagnostic standard based on the assessments of our experimental participants, (ii) using a continuous measure of disease likelihood that simply averages the assessments of the Mount Sinai labelers, (iii) restricting to cases where the diagnostic standard is definitive,[26] and (iv) a diagnostic standard that uses a lower threshold for determining a positive case

[26] Here, we restrict to cases where we can reject that the average assessment of the five Mount Sinai radiologists used to construct the diagnostic standard is equal to 0.5 at the 5% level (i.e., cases where we can reject the null hypothesis that $\sum_h \pi_h \left(\omega_i = 1 | s_{i,h}^H\right) / 5 = 0.5$).

(i.e. $\omega_i = 1\left[\sum_h \pi_h\left(\omega_i = 1|s^H_{i,h}\right)/5 > 0.3\right]$).

We also investigated the potential for mis-calibrated reports and experimental incentives biasing our results. The qualitative patterns of the treatment effects are unchanged if we calibrate each radiologists' assessments to the diagnostic standard before conducting the analysis. Incentives for accuracy, which are cross-randomized in designs 1 and 3, also do not have significantly different effects. Recall that our participants perform on par with the radiologists originally assigned to diagnose the patient cases.

Finally, this section treats pathologies are separable and does not account for potential interactions across pathologies for a given case. In section 3.5, we will present evidence showing that the model with the best fit has radiologists updating their beliefs as if pathologies are considered independently.

3.5 Automation Bias/Neglect and Signal-Dependence Neglect

An upshot of the results in section 3.4 is that our participants have valuable information, but they deviate from the benchmark of a Bayesian with correct beliefs about the joint distribution of their own information and the AI signal. These biases undercut the potential information advantage in a setup that involves AI assistance.

In this section, we theoretically model and estimate systematic deviations from this benchmark – which we will refer to as Bayesian for short – and determine the implications of these deviations for utilizing human expertise and AI predictions.[27] The next section empirically studies the optimal policy.

3.5.1 A Model of Deviations from Bayesian Updating

The framework in section 3.2 shows that a key question is whether the odds-ratios

$$\frac{p_h\left(\omega_i = 1|s^A_i, s^H_{ih}\right)}{p_h\left(\omega_i = 0|s^A_i, s^H_{ih}\right)} \text{ and } \frac{\pi_h\left(\omega_i = 1|s^A_i, s^H_{ih}\right)}{\pi_h\left(\omega_i = 0|s^A_i, s^H_{ih}\right)}$$

differ from each other. We now consider a set of models of belief-updating to describe systematic deviations from the Bayesian benchmark. In our model, the human correctly interprets their own signal when AI assistance is not available but errs when both s^A_i and s^H_{ih} are observed. As we will show below, whether or not AI assistance improves performance depends on the type of error humans make.

[27] The omission of the qualifier "with correct beliefs" slightly abuses terminology because a possible explanation of the deviations we have documented is that our participants are Bayesians but update their beliefs using an incorrect model for the joint distribution of s^A, s^H, and ω. We will entertain this possibility below.

The first class of biases that we consider arises when the two terms on the right-hand side of equation (3.5) are incorrectly weighted. Following Grether [1980, 1992], we parametrize this type of error using the following parsimonious functional form:

$$\log \frac{p_h\left(\omega_i = 1|s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0|s_i^A, s_{ih}^H\right)} = b_h \log \frac{\pi_h\left(s_i^A|\omega_i = 1, s_{ih}^H\right)}{\pi_h\left(s_i^A|\omega_i = 0, s_{ih}^H\right)} + d_h \log \frac{\pi_h\left(\omega_i = 1|s_{ih}^H\right)}{\pi_h\left(\omega_i = 0|s_{ih}^H\right)}, \tag{3.5}$$

where $b_h, d_h \geq 0$. The Bayesian is a special case with $b_h = d_h = 1$. While this linear form is restrictive, it has been useful for documenting several empirical regularities showing deviations from Bayesian updating, like base-rate neglect and under inference [see Benjamin, 2019, for a review].

We will say that the human exhibits *automation bias* if $b_h > d_h$ and *automation neglect* if $b_h < d_h$. As a motivation for this nomenclature, observe that when $b_h > d_h$, the human over-weights the AI signal relative to their own. Our theoretical analysis will focus on the case when $d_h = 1$, which is the empirically relevant case. The agent overshoots when updating the posterior odds relative to a Bayesian. Analogously, if $b_h < d_h$, then the human under-weights the AI signal relative to their own.[28]

A second class of deviations we consider will allow for models in which decision-makers do not account for the dependence between s_i^A and s_{ih}^H, which we call *signal dependence neglect*. For example, if humans act as if s_i^A and s_{ih}^H are independent conditional on ω_i even if they are not, then their posterior beliefs can be written as

$$\log \frac{p_h\left(\omega_i = 1|s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0|s_i^A, s_{ih}^H\right)} = b_h \log \frac{\pi_h\left(s_i^A|\omega_i = 1\right)}{\pi_h\left(s_i^A|\omega_i = 0\right)} + d_h \log \frac{\pi_h\left(\omega_i = 1|s_{ih}^H\right)}{\pi_h\left(\omega_i = 0|s_{ih}^H\right)}, \tag{3.6}$$

where b_h and d_h are allowed to differ from 1 as above. In the case when the signals are jointly multivariate normal and $b_h = d_h = 1$, signal dependence neglect yields correlation neglect as defined in [Enke and Zimmermann, 2019].[29] More generally, we will consider models that vary the conditioning set in the first term on the right-hand side and the dimension of s_i^A in the first term on the right-hand side. The specific examples are motivated and discussed further in section 3.5.3 below.

We intend for the models above to capture "as if" descriptions of humans' updating rules

[28] It is conceptually possible for d_h to differ from 1, which are similar in spirit to base-rate biases but apply to beliefs given the expert's signals instead of unconditional population rates [see Griffin and Tversky, 1992, Kahneman and Tversky, 1973]. This case is theoretically analyzed in a prior working paper version. Details are available on request.

[29] If s_i^A and s_{ih}^H are unidimensional with $\left(s_i^A, s_{ih}^H\right) \sim N\left(0, \Sigma_h\right)$, then the covariance matrix Σ_h is a sufficient statistic for the posterior probability that $\omega_i = 1$ given the signals if $\omega_i = 1\left\{s_i^A + s_{ih}^H \geq \varepsilon_i\right\}$ and ε_i is independent of $\left(s_i^A, s_{ih}^H\right)$.

and will remain agnostic about underlying mechanisms and micro-foundations. In particular, we remain silent on whether our participants are Bayesians who are utilizing the incorrect joint distribution of $\left(\omega_i, s_i^A, s_{ih}^H\right)$ when updating their beliefs or if they are non-Bayesians. The former type of model, known as a quasi-Bayesian model,[30] can generate automation bias or neglect as well as correlation biases.[31] An implicit assumption in our model, and likely other micro-foundations for the functional forms above as well, is that the signal acquired by the human is invariant to the provision of AI assistance. Whether additional training or experience with the AI can correct deviations from the benchmark model is therefore something that we leave for future work.

Nonetheless, the models above will prove useful for our purposes. From a theoretical perspective, the models will help outline the types of deviations that potentially decrease decision quality. From an empirical perspective, the models help clarify the drivers of the treatment effects documented earlier and turn out to be a good approximation to the data from the experiment.

3.5.2 Implications for Human-AI Collaboration

We now show that the types of deviations described above have implications for when AI assistance unambiguously improves human performance. The results will also illustrate the benefit of the simple functional forms in equations (3.5) and (3.6). This subsection drops the i and h indices for simplicity of notation.

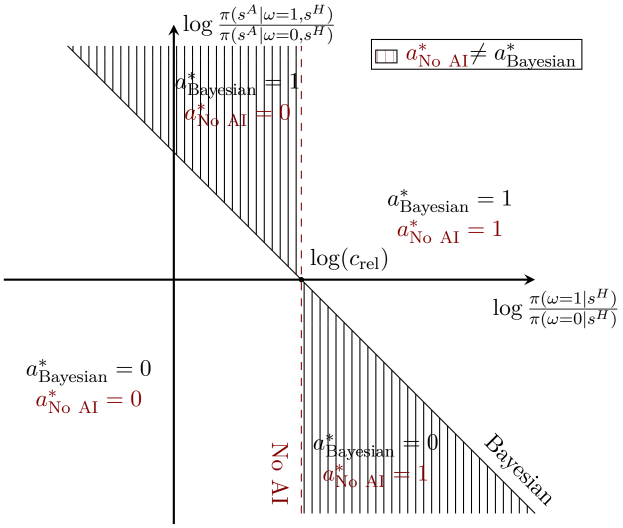
It is useful to start by considering the decisions with and without AI assistance for a Bayesian decision-maker. Figure 3.6 illustrates the realizations of s^A for which the optimal decision with AI assistance differs from the the decision without AI assistance for a fixed c_{rel}. The horizontal and vertical axes respectively represent $\log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ and $\log \frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$. As shown by the vertical dashed line, the decision-maker would take action 1 if and only if $\log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ exceeds $\log c_{rel}$. The solid line represents the analogous boundary for a Bayesian who has access to AI assistance. Observe that the decisions a Bayesian makes as a function of the signals s^A and s^H cannot be improved without additional information. Thus, a Bayesian and access to both signals improves upon the no-AI action in the vertically shaded region.

Now consider humans who may deviate from this benchmark model. A human who takes a given action without AI assistance $a^*_{\text{No AI}} = a^*\left(s^H; p_h\right)$ but a different action with AI

[30] See Rabin [2013] for a definition and Barberis et al. [1998], Rabin [2002], Rabin and Vayanos [2010] for examples.

[31] To see this, assume that $p_h\left(s_i^A|\omega_i, s_{ih}^H\right) = \pi\left(s_i^A|\omega_i, s_{ih}^H\right)^b$ and $p_h\left(s_{ih}^H, \omega_i\right) = \pi_h\left(s_i^H, \omega_i\right)$ to generate the functional form in equation (3.5) for any b_h as long as $d_h = 1$. The derivation of equation (3.6) is similar. In contrast to automation bias/neglect and correlation biases, own-information bias/neglect cannot be derived in a quasi-Bayesian model because we assume that $p_h\left(\omega_i|s_{ih}^H\right) = \pi_h\left(\omega_i|s_{ih}^H\right)$.

Figure 3.6: Comparing decisions with and without AI assistance – Bayesian



$\log \frac{\pi(s^A|\omega=1,s^H)}{\pi(s^A|\omega=0,s^H)}$
$a^*_{\text{No AI}} \neq a^*_{\text{Bayesian}}$
$a^*_{\text{Bayesian}} = 1$
$a^*_{\text{No AI}} = 0$
$a^*_{\text{Bayesian}} = 1$
$a^*_{\text{No AI}} = 1$
$\log(c_{\text{rel}})$
$\log \frac{\pi(\omega=1|s^H)}{\pi(\omega=0|s^H)}$
$a^*_{\text{Bayesian}} = 0$
$a^*_{\text{No AI}} = 0$
No AI
$a^*_{\text{Bayesian}} = 0$
$a^*_{\text{No AI}} = 1$
Bayesian

Note: Decision criterion of a Bayesian with and without AI assistance and where their decisions align. Shaded regions show the regions in which AI improves or worsens decision making.

assistance (so that $a^*_{\text{No AI}} \neq a^*_{\text{AI}}$) makes a worse decision if $a^*_{\text{AI}} = a^*\left(s^A, s^H; p_h\right)$ disagrees with a Bayesian's decision $a^*_{\text{Bayesian}} = a^*\left(s^A, s^H; \pi_h\right)$ with AI assistance. This follows because, in the binary action setup, only one of the decisions can agree with the Bayesian decision. In all other cases, the human's decision is weakly improved for the signal realization $\left(s^A, s^H\right)$. In other words, a human whose decision changes upon receiving the AI signal s^A is better off with AI assistance only if the change agrees with the Bayesian decision. The human is unambiguously better off if this property holds for all signals.

Our first result states that a human who exhibits automation neglect and no other deviation from the Bayesian model is unambiguously better off with AI assistance.

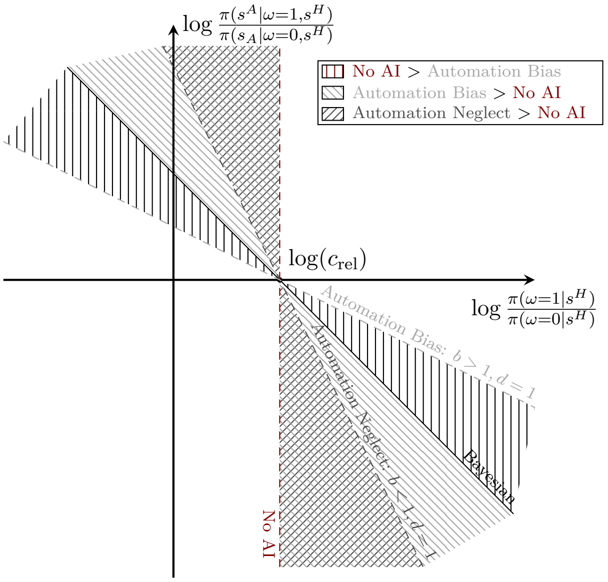
Proposition 3.5.1. *Suppose that the human's posterior is described by equation (3.5) and $d = 1$.*

(i) If the human exhibits automation neglect ($b < d = 1$), then for all pairs of signal realizations $\left(s^A, s^H\right)$, and any c_{rel}, the human attains weakly higher expected payoff $V(s)$ with AI assistance.

(ii) If the human exhibits automation bias ($b > d = 1$), for any c_{rel}, there exist log-likelihood ratios $\log \frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}$ and $\log \frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ such that the human attains lower expected payoff $V(s)$ with AI assistance.

See appendix C.1 for the proof.

Figure 3.7: Automation bias and neglect



Note: Where the decisions of an expert as a function of the signals disagree with a Bayesian in cases with and without AI assistance in the presence of automation bias or neglect when $d = 1$.

Figure 3.7 illustrates the result. The two dashed lines represent cutoffs analogous to those in figure 3.6 for humans with automation bias and automation neglect. Although a human who only exhibits automation neglect under-responds to the AI information, their beliefs move towards those of a Bayesian decision-maker but do not overshoot them. Whenever their decision changes, it agrees with the Bayesian's. In contrast, if the human exhibits automation bias, they err for moderately informative AI signals with intermediate values of $\log \frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$ because they over-react. At high enough values of this log-likelihood ratio, both the Bayesian and the human exhibiting automation bias would take the same action.

We next consider a decision-maker with a different type of bias, namely, one in which the decision-maker exhibits signal dependence neglect. Perhaps not surprisingly, our next result shows that this type of bias on its own can result in worse decisions with AI assistance:

Proposition 3.5.2. *Suppose that the human exhibits signal dependence neglect so that the posterior belief is described by equation (3.6). For any value of $b > 0$, $d > 0$, and $c_{rel} > 0$, there exist log-likelihood ratios $\log \frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$ and $\log \frac{\pi\left(s^A|\omega=1\right)}{\pi(s^A|\omega=0)}$ such that the human attains lower expected payoff $V(s)$ with AI assistance.*

See appendix C.1 for the proof.

Thus, signal dependence neglect adds another dimension of potential mistakes to those illustrated in the figures above. In its presence, there may be a joint distribution of signals for which AI assistance reduces performance. Even when $b = d = 1$ so that automation bias/neglect are not relevant, an examination of equations (3.5) and (3.6) reveals that whether or not a decision-maker exhibits under- or over-updating depends on the difference between $\log \frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$ and $\log \frac{\pi\left(s^A|\omega=1\right)}{\pi(s^A|\omega=0)}$.

The result bears resemblance to those in Enke and Zimmermann [2019], which shows that in a multivariate normal model with positively correlated signals, correlation neglect results in over-reaction to signals and verified this hypothesis in lab experiments. Proposition 3.5.2 differs in that it allows for general signal distributions and for signal dependence neglect to coexist with automation bias/neglect. This extension is essential for a naturalistic environment like ours because the experimenter does not have full control over the signal structure. The general signal structure makes it difficult to characterize mistakes in terms of over or under-updating, unlike in the case of a multivariate normal model.

The propositions above have important implications for the design of human-AI collaboration, which we consider in section 3.6. Specifically, we study an AI designer who only has access to the AI signal s^A and must decide on one of the three modes of delegation: utilize only the AI prediction, delegate the case to the human, or provide AI assistance to a human expert. The results show that other than in the case when automation neglect is the only relevant bias, the designer must learn the types of biases as well as the distribution of $\pi\left(s^A, s^H|\omega\right)$ to determine which delegation modality yields the best decision.

3.5.3 Estimating Deviations from Bayesian Updating

We now turn to an empirical implementation of the model above. The analysis in this section will be based on designs 2 and 3 because they allow us to observe the same participant make decisions under all information-conditions on a given case. Consider the empirical analog to equation (3.5):

$$\log \frac{p_h\left(\omega_i = 1|s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0|s_i^A, s_{ih}^H\right)} = a + b \cdot \log \frac{\pi_h\left(s_i^A|\omega_i = 1, s_{ih}^H\right)}{\pi_h\left(s_i^A|\omega_i = 0, s_{ih}^H\right)} + d \cdot \log \frac{\pi_h\left(\omega_i = 1|s_{ih}^H\right)}{\pi_h\left(\omega_i = 0|s_{ih}^H\right)} + \varepsilon_{ih}, \quad (3.7)$$

where we have omitted heterogeneity across radiologists in b_h and d_h (appendix C.3.6 discusses radiologist heterogeneity in these estimates). Two of the terms in this equation are directly elicited: the probability in the second term on the right-hand side, $\pi_h(\omega_i = 1|s_{ih}^H)$, is set to the radiologists' assessment without AI assistance and the term $p_h\left(\omega = 1 \left| s_{ih}^A, s_{ih}^H\right.\right)$

in the dependent variable is the assessment in the treatment arm with AI.[32] The "update term," given by $\log \frac{\pi_h\left(s_{ih}^A|\omega_i=1,s_{ih}^H\right)}{\pi_h\left(s_{ih}^A|\omega_i=0,s_{ih}^H\right)}$ will be estimated and substituted into the equation above.

There are three challenges in estimating the update term. The first challenge is that it is a ratio of conditional densities. We address this issue by rewriting it using Bayes' rule as follows:

$$\log \frac{\pi_h\left(s_i^A|\omega_i=1,s_{ih}^H\right)}{\pi_h\left(s_i^A|\omega_i=0,s_{ih}^H\right)} = \log \frac{\pi_h\left(\omega_i=1|s_i^A,s_{ih}^H\right)}{\pi_h\left(\omega_i=0|s_i^A,s_{ih}^H\right)} - \log \frac{\pi_h\left(\omega_i=1|s_{ih}^H\right)}{\pi_h\left(\omega_i=0|s_{ih}^H\right)}.$$

If s_{ih}^H can be constructed or controlled for, then we can estimate the first term on the right-hand side using data on ω_i and s_{ih}^A via a binary response model. Observing s_i^A is immediate because the signal from the AI given to humans is isomorphic to the vector of predicted probabilities for the various pathologies. The second term in this equation has been elicited.

This brings us to the second challenge, which is controlling for s_{ih}^H when estimating $\pi_h\left(\omega_i=1|s_i^A,s_{ih}^H\right)$ because we do not observe it directly, unlike in a laboratory setting (c.f. Conlon et al., 2022). If s_{ih}^H is unidimensional and $\pi_h\left(\omega_i|\, s_{ih}^H\right)$ is monotonic in s_{ih}^H, then $\pi_h\left(\omega_i|\, s_{ih}^H\right)$ is a valid control variable. However, we want to allow for the possibility that the radiologist evaluates a case holistically and uses signals across pathologies. Our empirical specifications will therefore employ multivariate proxy controls for s_{ih}^H using elicited probability assessments for multiple pathologies.[33]

To allow for flexible interactions between s_i^A and s_{ih}^H while avoiding over-fitting, we estimate $\pi_h\left(\omega_i=1|s_i^A,s_{ih}^H\right)$ using a pathology-specific random forest that predicts ω_i using the vector of predicted probabilities for all pathologies for case c_i reported by radiologist h without AI assistance, the vector of predicted probabilities for case c_i the AI algorithm produces, summaries of the patient clinical history when made available to the radiologist, and participant-specific fixed-effects.[34]

The third challenge is the potential for measurement error, particularly of the form that radiologists' signal s_{ih}^H when elicited without AI might differ from their signal when given AI assistance. Classical measurement error arising from this source would lead to attenuation bias in the coefficient estimates. To address this issue, we will construct instruments for s_{ih}^H

[32] To avoid undefined terms in calculating log-odds ratios we take the minimum of all probability assessments and 0.95 and the maximum of all probability assessments and 0.05.

[33] Specifically, we will use the vector of probabilities for all pathologies reported by h for case i, $\left(\pi_h\left(\omega_{i'}|\, s_{i'h}^H\right)\right)_{i'\in I(c_i)}$, as the control variable. Here, c_i is the patient case associated with case-pathology i and $I\left(c_i\right)$ is the set of case-pathologies considered when deciding case c_i. This control variable is valid under the assumption that $s_i^A \perp s_{ih}^H\big|\, \omega_i, \left(\pi_h\left(\omega_{i'}|\, s_{i'h}^H\right)\right)_{i'\in I(c_i)}$.

[34] The hyper-parameters of the random forest are chosen by grouped k-fold cross-validation, where we ensure that each patient case appears in only one fold to avoid overfitting to the patient case. Further details of the training procedure are described in appendix C.3.6.

using the reported probabilities of the other radiologists in our experiment.

With these solutions in hand, we would like to assess whether humans exhibit signal dependence neglect when updating beliefs. As prefaced earlier, although the human and AI signals are not conditionally independent given the diagnostic standard, humans may act as if they are. We will therefore estimate and select between models that vary the set of signals conditioned on in the update term. For example, in the case when radiologists behave as if s_i^A and s_{ih}^H are independent conditional on ω_i, the update term in equation (3.7) drops the conditioning on s_{ih}^H.[35]

The correct model of behavior satisfies the moment restriction $E\left[\varepsilon_{iht}\middle|\, s_{i,-h}^H, s_i^A\right] = 0$, where $s_{i,-h}^H$ collects the signals of the radiologists other than h in our experiment. For estimation, we utilize unconditional moment restrictions based on functions of $s_{i,-h}^H$ and s_i^A that closely mimic the terms in equation (3.7). Our instruments include $\log \frac{\pi\left(\omega_i=1|s_i^A\right)}{\pi\left(\omega_i=0|s_i^A\right)}$ and leave-one-out averages of $\log \frac{\pi\left(\omega_i=1|s_i^A, s_{ih'}^H\right)}{\pi\left(\omega_i=0|s_i^A, s_{ih'}^H\right)}$ for radiologists other than h that use various proxies for s_{ih}^H using assessments from different sets of pathologies obtained without AI assistance.[36] Empirical analogs of the resulting moment conditions are used to estimate the model using GMM.

We will employ the model-selection procedure proposed in Andrews and Lu [2001] to select between non-nested models. The method uses a selection criterion, the MMSC-BIC, which is constructed from the J-statistic of the GMM objective function with an aditional term that penalizes models that reject a greater number of moment restrictions.

3.5.4 Results

Our results indicate that while there are large potential gains from combining radiologists' assessments with AI predictions, biases in radiologists' use of AI assistance undercuts these gains. We find that radiologists exhibit both automation neglect and signal dependence neglect. These mistakes prevent AI assistance from improving diagnostic performance.

Table 3.2 presents estimates from six "as if" models of participant behavior.[37] According

[35] We can vary the pathologies across the set of models considered when constructing $I\left(c_i\right)$. The conditionally independent case corresponds to the extreme case in which $I\left(c_i\right) = \emptyset$, whereas the Bayesian model includes all pathologies.

[36] Specifically, we construct 14 instruments. The first is a constant and the second is the average of $\log \frac{\pi\left(\omega_i=1|s_{ih'}^H\right)}{\pi\left(\omega_i=0|s_{ih'}^H\right)}$ for all $h' \neq h$. The remaining 12 construct the average of $\log \frac{\pi(\omega_i=1|s_{ih'})}{\pi(\omega_i=0|s_{ih'})}$ for all $h' \neq h$ by varying the conditioning variables s_{ih}. The different sets of conditioning variables in s_{ih} are presented in the second panel of appendix table C.26. These sets are used because they are the relevant terms in at least one of the models that we consider in the testing procedure.

[37] These are a subset of the full set that we consider. See table C.26 in the appendix for the results from all models. Results from all pathologies with AI are qualitatively similar (see table C.27). These analyses were pre-registered except for the model selection exercise, which was not included in the pre-analysis plan.

to the first model, participants act as if s_i^A and s_{ih}^H are conditionally independent given ω_i and consider each pathology separately. The second model accounts for dependence between s_i^A and s_{ih}^H but maintains the assumption that pathologies are considered separately. The third model accounts for dependence across pathologies in diagnosis by including signals from other pathologies in s_i^A and s_{ih}^H. The next three models are identical to the first three but include clinical history information (when provided) in s_{ih}^H. Setting $b = d = 1$ and the constant to 0 in the last model corresponds to Bayesian updating with correct beliefs.

The results from this exercise point to two types of errors in radiologists' use of AI signals. The first type of error is that radiologists neglect signal dependence even though AI predictions and radiologists' signals are highly correlated after conditioning on the diagnostic standard (see appendix table C.24). This conclusion follows because we select the model in column 1 as it has the lowest value of the MMSC-BIC statistic. Another implication of the selected model is that radiologists do not incorporate information across different pathologies since only the focal pathology is relevant. This result validates our previous analysis that evaluates each pathology separately. The second type of error is that radiologists exhibit automation neglect, and we estimate a value of d that is close to 1 across all models we consider.

Connecting these observations back to our theoretical discussion in section 3.2, the parameters are such that access to the AI signal may not improve performance, primarily because of signal dependence neglect.

Table 3.2: Selecting between models of belief updating: top level pathologies with AI

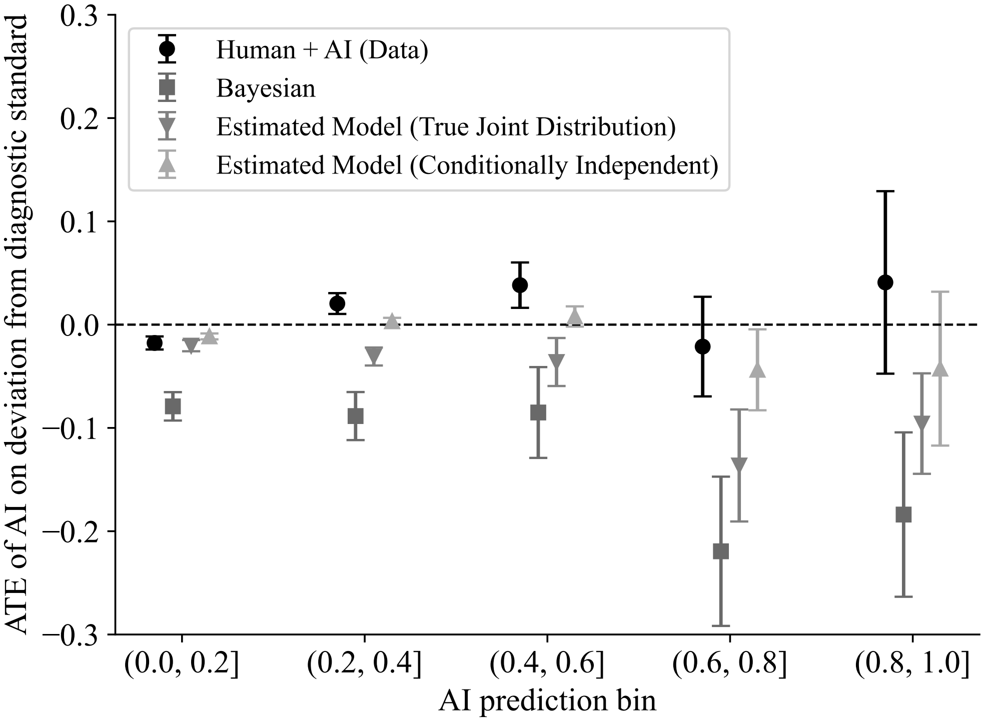
	(1)	(2)	(3)	(4)	(5)	(6)
Automation bias (b)	0.27	0.33	0.12	0.19	0.21	0.12
	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
Own information bias (d)	1.11	1.09	1.05	1.07	1.07	1.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.39	0.39	0.25	0.32	0.32	0.25
	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)
No signal dependence	✓			✓		
No pathology dependence	✓	✓		✓	✓	
No clinical history dependence	✓	✓	✓			
Correct updating						✓
J-Statistic	13.08	11.63	8.85	7.53	8.55	7.72
MMSC-BIC	-29.11	-26.34	-8.03	-13.57	-12.55	-9.16
Selected moments	13	12	7	8	8	7
Possible moments	14	14	14	14	14	14
Observations	11420	11420	11420	11420	11420	11420
R-Squared	0.49	0.48	0.42	0.45	0.45	0.42

Note: Estimates of b and d for different specifications of the update term. The models differ by whether the update term conditions on the signal s_H of the pathology at hand, the AI and the radiologist signals for other pathologies, and the information provided in the clinical history of a patient. Each model is estimated via GMM. The reported MMSC-BIC statistic adjusts the J-statistic for the number of included parameters and moments, awarding bonus terms for models with fewer parameters and fewer rejected moments (see Andrews and Lu [2001] for details). The full set of models in the selection procedure are presented in table C.26. The update term is estimated via random forest as described in appendix section C.3.6. Standard errors are clustered at the radiologist level. This table uses data from designs 2 and 3 where we observe the same human's assessment of each case both with and without AI assistance.

These deviations also explain the heterogeneous conditional average treatment effects documented in section 3.4. Figure 3.8 shows the estimated conditional average treatment effect of AI alongside model-implied treatment effects from three scenarios: a Bayesian benchmark, the model in column (6) where radiologists only exhibit automation neglect, and the selected model from table 3.2. As expected, a Bayesian performs significantly better when given the AI signal. In fact, as indicated by the large reductions in the deviation from the diagnostic standard, there is significant potential value in combining the human and AI signals. The model that only features automation neglect – column (6), equation (3.5) – reduces these improvements and moves the implied treatment effects closer to the data. However, throughout the entire signal range of AI predictions, the performance of such a decision-maker would still unambiguously increase with AI assistance, consistent with our theoretical model's prediction. Only when we use the selected model, under which radiologists neglect signal dependence, can we replicate the worsening of assessments with AI in the middle of the signal range.

Although the specifications above replicate the pattern of conditional treatment effects,

Figure 3.8: Data versus model implied treatment effects



Note: Observed conditional treatment effects of providing radiologists access to AI compared to three different model-implied treatment effects: giving AI access to a Bayesian decision-maker, giving AI access to a decision-maker who acts according to the empirical version of equation 3.5 both under the correct update term and when the decision-maker treats the AI signal as conditionally independent. Standard errors are two-way clustered at the radiologist and patient-case level, with 95% confidence intervals depicted. Standard errors on model based treatment effects are conditional on the model of behavior. We first generate $p_h\left(\omega_i = 1|s_i^A, s_i^H\right)$ based on the model in Equation 3.7 and then estimate the treatment effect and standard error of these model-based posteriors ($p_h\left(\omega_i = 1|s_i^A, s_i^H\right)$).

one may still be concerned about mis-specification of the model of belief-updating or about heterogeneity in b and d across radiologists. For example, a model of updating in which radiologists' beliefs move to the maximum of their own predictions and AI predictions would not be linear in log-odds. Appendix C.3.6 uses a non-parametric model to show that the relevant boundary of the decision-regions depicted in figure 3.7 is well approximated using the linear specifications considered. Appendix C.3.6 investigates heterogeneity by allowing b and d to vary with radiologists. We find that the estimated distributions of b_h and d_h are centered close to the point estimates above, with most of the estimated distributions of b_h and d_h in the range $[0.1, 0.4]$ and $[1.0, 1.2]$, respectively. Together, these results suggest the specification in column (1) of table 3.2 represents a good approximation to data we collected.

3.6 Designing Human-AI Collaboration

We now consider the design of collaborative systems between AI and humans. Because the AI signal can be obtained at zero marginal cost we consider a policy $\tau(\cdot)$ that chooses between full automation (AI), humans with access to AI ($H + AI$), or humans without access to AI (H), as a function of the AI signal s_i^A. We then compare this policy in terms of human time cost and decision loss to policies where all cases are exclusively decided by either the AI, humans, or humans with access to AI.

As a warm-up for this exercise, it is useful to examine the predictive performance of the different modalities, conditional on s^A (see figure 3.9). Recall from the conditional treatment effect analysis that human assessments improve with AI in the lowest and second-highest bins of AI signals. Figure 3.9 also shows that even when the AI improves human decision-making (in the lowest bin), AI alone outperforms humans with AI. Although this figure does not account for differences in the human time costs across modalities, our analysis of the estimated treatment effects shows that humans take more time when provided with AI predictions. This points to the conclusion that in most cases where AI improves decision-making, one is at least as well off relying exclusively on AI predictions because humans do not incoporate the information effectively and are slower when deciding with AI. However, automating all cases is not necessarily optimal either because humans perform better than AI when the AI is uncertain.

Motivated by these observations, we now examine if there is a trade-off between the marginal costs of human effort and diagnostic performance.

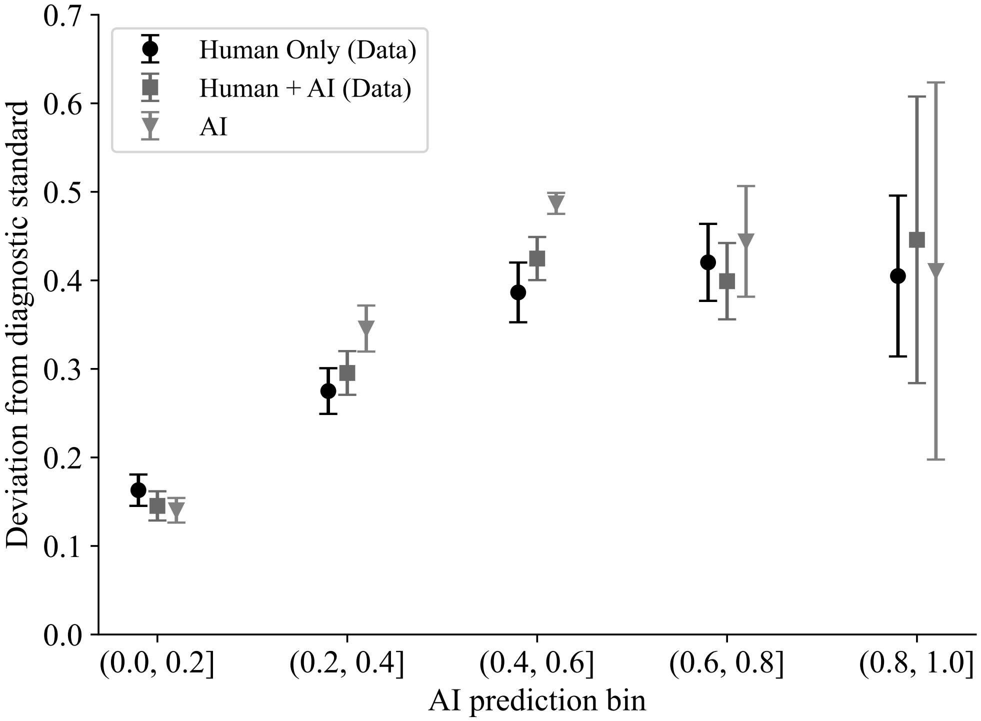
3.6.1 Computing the Trade-off Between Decision Loss and Costs of Human Effort

The optimal policy which minimizes the sum of the expected decision-loss (costs of false positives and false negatives) and the monetized time cost of using humans solves:

$$\tau^*(s_i^A) = \arg \min_{\tau \in \{H, H+AI, AI\}} m \cdot V_\tau \left(s_i^A\right) + w \cdot C_\tau \left(s_i^A\right). \tag{3.8}$$

The first term contains the expected decision-loss from a modality given by $V_\tau \left(s_i^A\right) = E\left[V_{ih\tau} \middle| s_i^A\right]$, which is the expected diagnostic quality given the AI signal. The expectation is taken over both cases and radiologists. The parameter m is the dollar cost of a false negative (e.g., a missed diagnosis). We allow preferences for false positives and false negatives to vary by pathology. We estimate these preferences using data on the binary treatment recommendations of the participants in our experiment, given their probability assessments.

Figure 3.9: Model deviation from diagnostic standard



Note: Performance of the different modalities that we consider for the optimal collaborative system. Cases are decided by either only the human, only the AI, or the human with access to the AI. The performance measures for Human Only and Human + AI are constructed from our treatment effect analysis. Standard errors are two-way clustered at the radiologist and patient-case level, with 95% confidence intervals depicted.

According to the model in section 3.2, human h's choice a_{hi} of recommending treatment or follow-up on patient case i is given by

$$a_{hi} = 1\left[\frac{p_{hi}}{1-p_{hi}} - c_{rel}^{h} + \varepsilon_{hi} > 0\right],$$

where p_{hi} is the human's belief about pathology presence, c_{rel}^{h} is the relative cost of false positives and false negatives, and ε_{hi} captures idiosyncratic unobserved preference heterogeneity. We allow the parameters of this model to vary by pathology, but we suppress this dependence for notational simplicity. The full set of results of this exercise are presented in Table C.32. The median cost of a false positive across both top-level pathologies with AI assistance is one half the cost of a false negative. Since we do not know the dollar cost of a false negative, we will present results for a range of values for m.

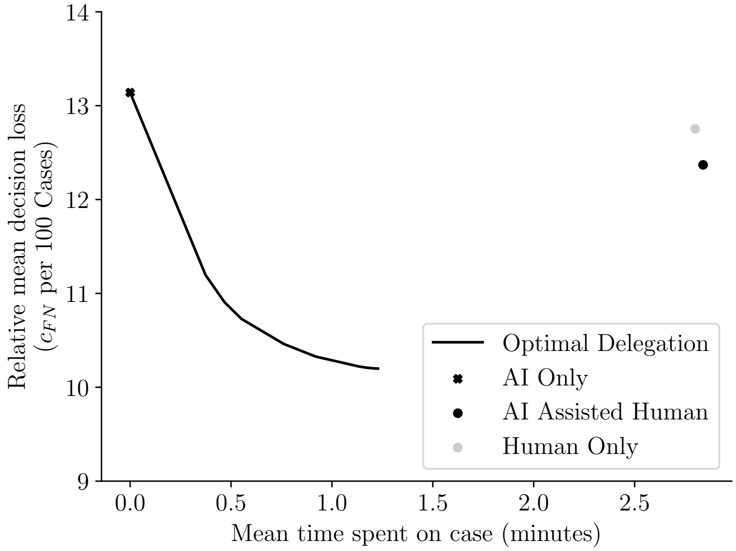
The second term in the objective function contains $C_{\tau}\left(s_i^A\right) = E\left[C_{ih\tau}\middle| s_i^A\right]$, which is the expected time cost for a given modality τ. If the case is fully automated, this time cost is zero. Otherwise, the time costs are based on our experimental estimates, which show that radiologists spend more time on cases when presented with AI predictions. For the costs of human radiologist time, we set $w = \$4$ per minute based on a payment of \$10 per case and the observed average time per read of approximately 2.5 minutes.

Next, we solve the problem in equation (3.8) by first estimating the conditional mean functions $V_\tau\left(s_i^A\right)$ and $C_\tau\left(s_i^A\right)$ using random forest regressions. We tune the random forest hyperparameters by using grouped cross-validation where observations are grouped by their patient case to avoid over-fitting to specific cases. Given estimates of expected diagnostic quality and the expected time cost for each modality on a given case, we assign the case to be read by the modality that minimizes Equation 3.8. We repeat this exercise for a range of values of m. In our discussion of the results we focus on airspace opacity but the qualitative findings remain unchanged if we consider other pathologies (appendix C.3.8).

3.6.2 Results

There are large potential gains from optimally delegating cases. Figure 3.10 shows a possibilities frontier for the trade-off between diagnostic quality against decision time, calculated by varying the social cost of false negatives m. One extreme on this frontier is the point where the AI decides all cases, thus minimizing the time costs. The figure shows that one can substantially reduce both time costs and decision loss by moving from H or $H + AI$ to the frontier. Table 3.3 provides an overview comparing H and $H + AI$ to the two extreme points of the frontier (i.e., AI only and the delegation policy that minimizes decision loss) along with a comparison to a Bayesian decision-maker. For each of those, the table compares the expected decision loss, the time cost (in minutes and dollars), and the fraction of false positives/negatives. An unassisted radiologist (H) takes 2.8 minutes minutes per case, or about \$11, and incurs a relative decision loss of approximately 12.8. By moving to the frontier point that minimizes decision loss, one can reduce decision loss while also saving \$6.3 in time costs. Similar gains can be achieved from $H + AI$. A Bayesian decision-maker incurs the lowest decision loss but faces the same time costs as $H + AI$.

Figure 3.10: Loss-time frontier



Note: Human radiologists and AI performance relative to the optimal delegation system on the frontier of the cost of human time versus decision loss. This analysis excludes data from design 3 because of learning effects in this setup.

Table 3.3: Airspace opacity delegation results

	Time Cost		Pr(Fp)	Pr(Fn)	Decision Loss
	Minutes	Dollars			
Bayesian	2.8	11.4	6.4	1.6	4.1
AI Only	0.0	0.0	32.4	1.2	13.5
Human Only	2.8	11.2	17.0	6.3	12.8
Human + AI	2.8	11.4	21.6	4.2	12.4
Min. Decision Loss	1.2	4.9	12.7	3.4	10.2

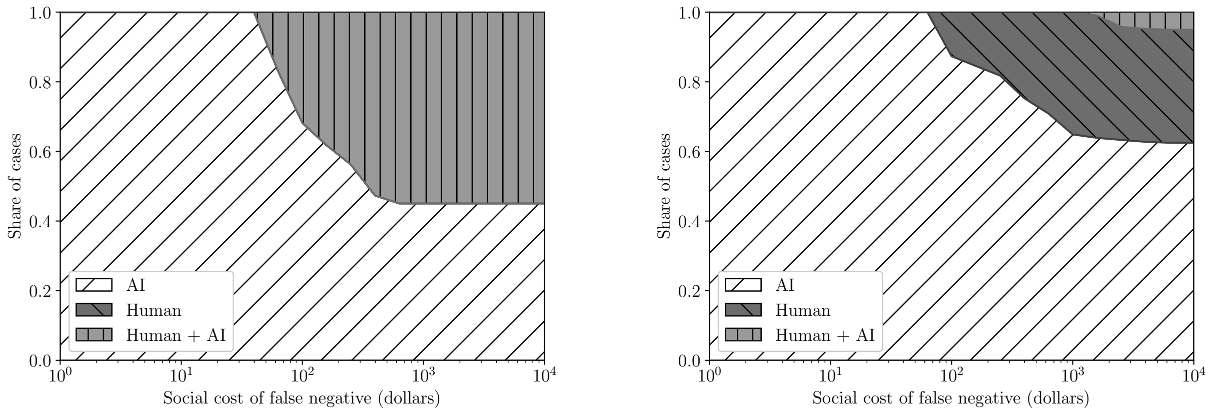
Note: Time taken and decision loss of delegation strategies for Airspace Opacity. The average time per case is shown in both minutes and dollars using a wage of $4 per minute. The table also reports the share of false positives ($Pr(FP)$), the share of false negatives ($Pr(FN)$), and decision loss calculated as $Pr(FN) + c_{rel}Pr(FP)$ where c_{rel} =0.38 – the median c_{rel} for Airspace Opacity. The Bayesian row shows results for the Bayesian decision-maker. AI Only shows results for full delegation to the AI. Human Only shows results if humans read cases without AI assistance. Human + AI shows results if humans with access to the AI read all cases. Min. Decision Loss shows results for the optimal delegation strategy that minimizes decision loss and highlights the potential improvement in decisions from delegating to the AI. This analysis excludes data from design 3 because of learning effects.

Next, we investigate what share of cases is decided by the three modalities under the optimal delegation policy as we vary m (figure 3.11). For both a Bayesian and the observed behavior in our experiment, we find that the AI decides almost all cases if the cost of a false negative is less than $100 per case. For Bayesians, the share of cases that involve human-AI collaboration rises markedly above a cost of $100, but even for costs as high as $10,000, 45% of cases are delegated to the AI. Moreover, under Bayesian decision-making, the share of cases where only the human decides the case without access to the AI signal is negligible and the only reason for using an unassisted human in this case is to save on time costs. When we conduct the same exercise and use the observed behavior of human radiologists, we find that humans are involved in 38% of cases if the cost of a false positive is sufficiently large. Moreover, the majority of cases where a human is involved have the human make decisions without AI assistance. A more complete assessment of the optimal combination of human and machine decisions, therefore, confirms the intuition from above that cases are either decided by humans or the AI but not by both of them together.

Figure 3.11: Airspace opacity modality shares

(a) Bayesian

(b) Humans



Note: Share of cases decided by each modality (humans, AI, humans+AI) conditional on the cost of a false negative in dollars, denoted m in the text, for airspace opacity. Panel (a) focuses on a Bayesian decision-maker. Panel (b) focuses on a human decision-maker with decisions and time taken as in our experiment. This analysis excludes data from design 3 because of learning effects.

3.6.3 Caveats

There are several caveats to our analysis. The first is that we consider AI assistance for a single pathology at a time. This approach abstracts away from interactions between pathologies. It is best suited to contexts in which the focal pathology of interest for a case is clear to a treating physician. Given our results in section 3.5, it appears that physicians do not account for cross-pathology interactions, thereby complicating attempts to infer such interactions from radiologist assessments and behavior. The second caveat is that any collaborative system may change humans' expectations about the difficulty of cases and adjust strategically to those changing expectations. Our approach abstracts away from endogenous information acquisition, for example, rational inattention as in Sims [2003]. A potentially interesting aspect is whether a designer can leverage such endogenous responses by designing an information revelation policy that induces effort. We leave such extensions that leverage insights from information design [Bergemann and Morris, 2019, Kamenica and Gentzkow, 2011] to future work, but we do measure the total amount of time taken by the experts in our experiment with and without AI assistance.

3.7 Conclusion

AI is predicted to profoundly reshape the nature of work [see Felten et al., 2023]. Humans are likely to use AI as a decision aid for many tasks not only in the long run but also in the medium run for tasks that will ultimately be fully automated. A central question is therefore how humans use AI tools and how tasks should be assigned. Radiology is an iconic example, one that employs a large number of professionals whose main job is a high-stakes classification task.

To understand the benefits and pitfalls of human-machine collaboration, we conduct an experiment in which AI assistance for radiologists is randomized. We also randomize the availability of contextual information that is typically available to radiologists but is not used to train AI prediction tools for chest X-rays. Since we can simulate radiologists' normal workflow, this is an ideal setting for conducting such an experiment. We then devise a methodology to estimate the radiologists' deviation from Bayesian updating, an approach which needs to deal with the challenge that we do not directly control the information structure that radiologists face when making decisions.

While deploying AI assistance in our setting has large potential benefits, biases in humans' use of AI assistance eliminate these gains. Even though the AI tool in our experiment performs better than two-thirds of radiologists, we find that giving radiologists access to AI predictions does not, on average, lead to higher performance. This average treatment effect, however, masks systematic heterogeneity: providing AI does improve radiologists' predictions and decisions for cases where the AI is certain (e.g., predicted probability is close to zero or one) but not when it is uncertain. This latter result – that prediction quality can be reduced for some range of AI signals – rejects Bayesian updating We also identify systematic errors in belief updating; specifically radiologists exhibit automation neglect (e.g., radiologists underweight the AI prediction relative to their own) and treat the AI prediction and their own signals as independent conditional on the correct class even though they are not. Moreover, radiologists take significantly more time to make a decision when AI information is provided.

Together, these results have important implications for how to design collaborations between humans and machines. Increased time costs and sub-optimal use of the AI information both work against having radiologists make decisions with AI assistance. In fact, an optimal delegation policy that utilizes heterogeneity in treatment effects given the AI prediction suggests that cases should either be decided by the AI alone or by the radiologist alone. Only a small share of cases are optimally delegated to radiologists with access to AI. In other words, we find that radiologists should work *next to* as opposed to *with* AI. To the extent that expert decision-makers generally under-respond to information other than their own [Conlon et al., 2022] and incorporating additional information is cognitively costly, these insights may hold in other settings where experts' main job is a classification task.

There are several important considerations that are outside the scope of this work. One question motivated by the unrealized potential gains of AI assistance concerns the benefits from AI-specific training for radiologists and/or experience with AI. This and related questions require different experimental designs or longer-run studies. Other open questions are whether the heterogeneity in the value of AI assistance is correlated with a human's baseline skill or other characteristics and whether such correlation can be predicted to target assistance. The organization of human-AI collaboration also raises questions about whether the form of collaboration influences humans' incentives to respond strategically. The use of AI in practice will also be mediated by other organizational incentives and the regulatory environment. Organizations may set guidelines on how to use AI or provide feedback, and regulations may influence liability implications. These issues are interesting avenues for future work.

AI continues to evolve rapidly. While economists are unlikely to have a major role in the technical development of AI tools, our comparative advantage lies in studying how humans interact with these tools and thereby helping shape the institutions that guide their use to ensure that this development is beneficial to society. Empirical analysis is a particularly useful tool in this endeavor, especially if the algorithms themselves are a black-box and cannot be understood from first principles.

Appendix A

Appendix: Personalized Rankings and User Engagement: An Empirical Evaluation of the Reddit News Feed

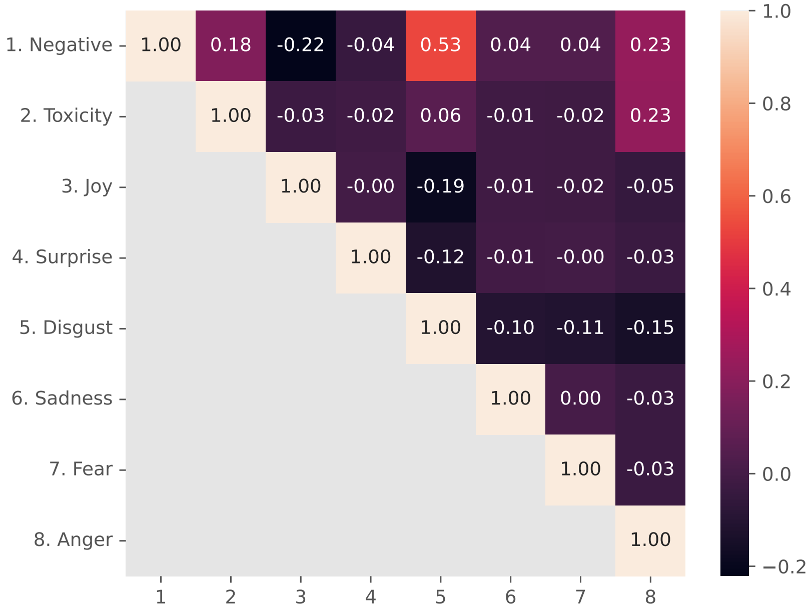
A.1 Data Appendix

Table A.1: Summary of Publisher Ratings

	Slant Score	Credibility
msnbc.com	-0.62	0.59
huffpost.com	-0.31	0.57
nytimes.com	-0.26	0.86
wsj.com	0.01	0.80
foxnews.com	0.61	0.53
breitbart.com	0.74	0.30

Note: Publisher slant and credibility ratings for six widely known publishers.

Figure A.1: Correlation Matrix of Text Features



Note: This figure plots the correlation matrix of comment text features. Negative corresponds to negative sentiment, Toxicity corresponds to the predicted toxicity of the comment, while the remaining 6 features correspond to the emotional content of the post.

A.2 Reduced Form Appendix

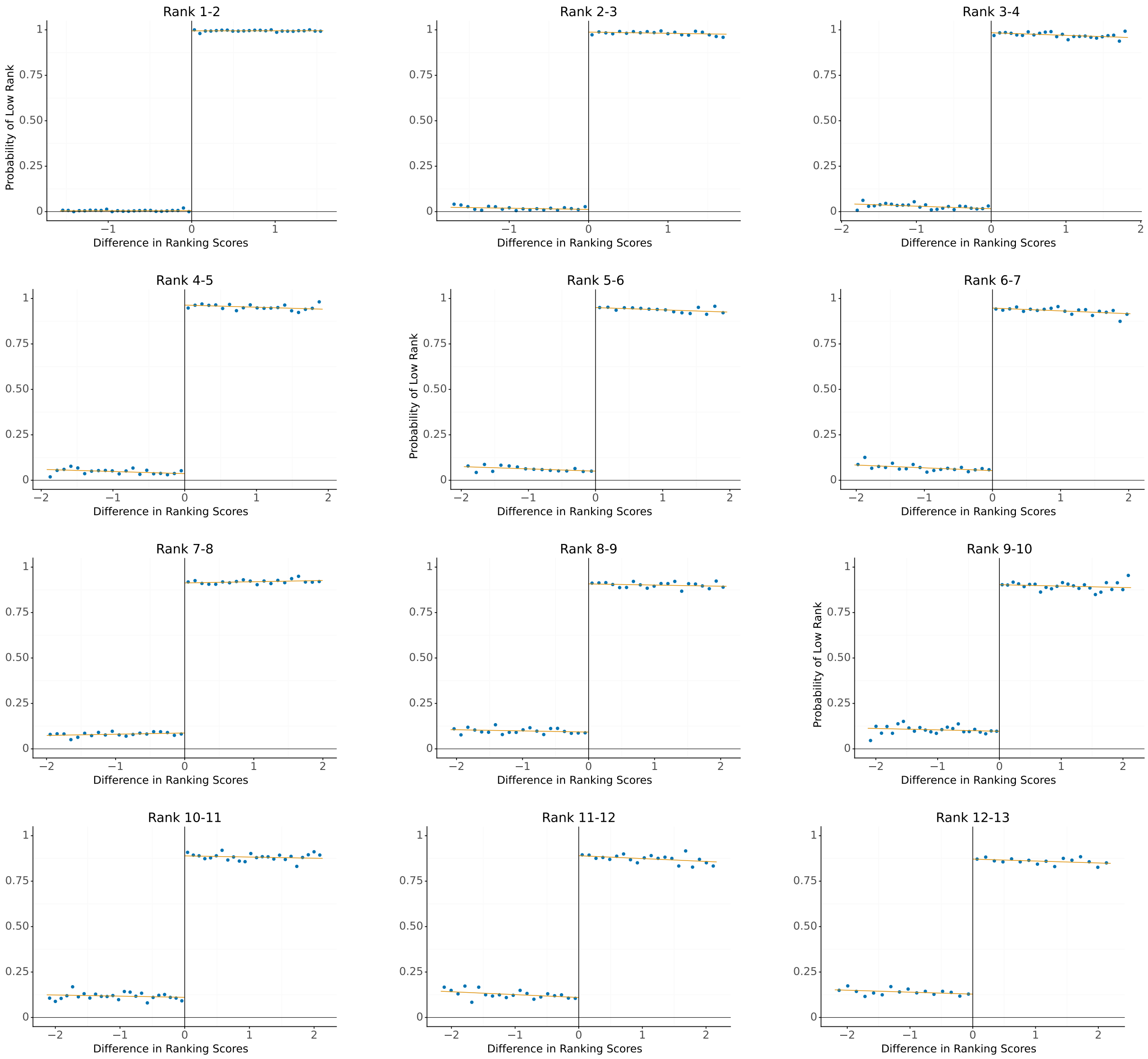
A.2.1 Additional Figures and Tables

Table A.2: Position Effect Estimates

		Regression Discontinuity		
Rank	OLS	Local Linear	Local Constant	Triangular Kernel
13	0.022	0.014	-0.007	0.019
	(0.007)	(0.029)	(0.061)	(0.032)
14	0.037	0.033	0.015	0.026
	(0.006)	(0.027)	(0.057)	(0.029)
15	0.022	0.049	0.082	0.045
	(0.006)	(0.023)	(0.047)	(0.025)
16	0.025	0.033	0.042	0.032
	(0.006)	(0.020)	(0.040)	(0.022)
17	0.018	0.024	0.054	0.022
	(0.006)	(0.022)	(0.045)	(0.024)
18	0.015	0.018	0.029	0.021
	(0.006)	(0.022)	(0.047)	(0.024)
19	0.023	0.006	0.001	0.017
	(0.005)	(0.023)	(0.049)	(0.026)
20	-0.001	-0.035	-0.081	-0.037
	(0.005)	(0.019)	(0.040)	(0.021)
21	0.018	0.024	0.018	0.019
	(0.005)	(0.019)	(0.041)	(0.021)
22	-0.002	-0.007	-0.000	-0.010
	(0.005)	(0.020)	(0.042)	(0.022)
23	0.010	0.020	0.040	0.007
	(0.005)	(0.022)	(0.047)	(0.025)
24	0.016	0.015	0.032	0.024
	(0.005)	(0.023)	(0.049)	(0.025)

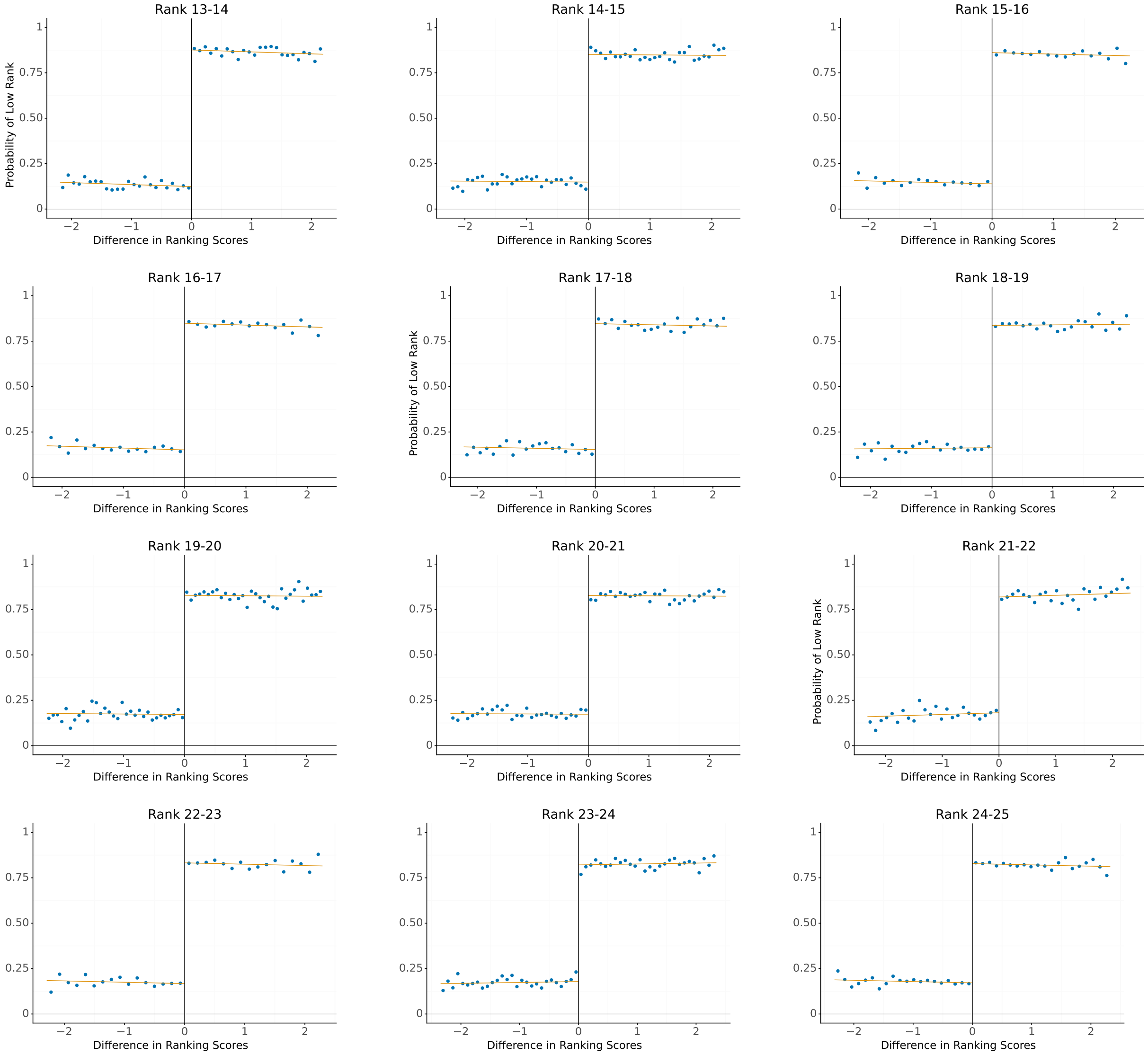
Note: Estimates of the local average treatment effect from a post moving from position $r + 1$ to position r on the feed on the log-number of comments a post receives. Robust bias-corrected standard errors that allow for misspecification of the conditional expectation function and that are clustered at the period level are shown in parenthesis. Estimates exclude posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05.

Figure A.2: Regression Discontinuity Plots: First Stage



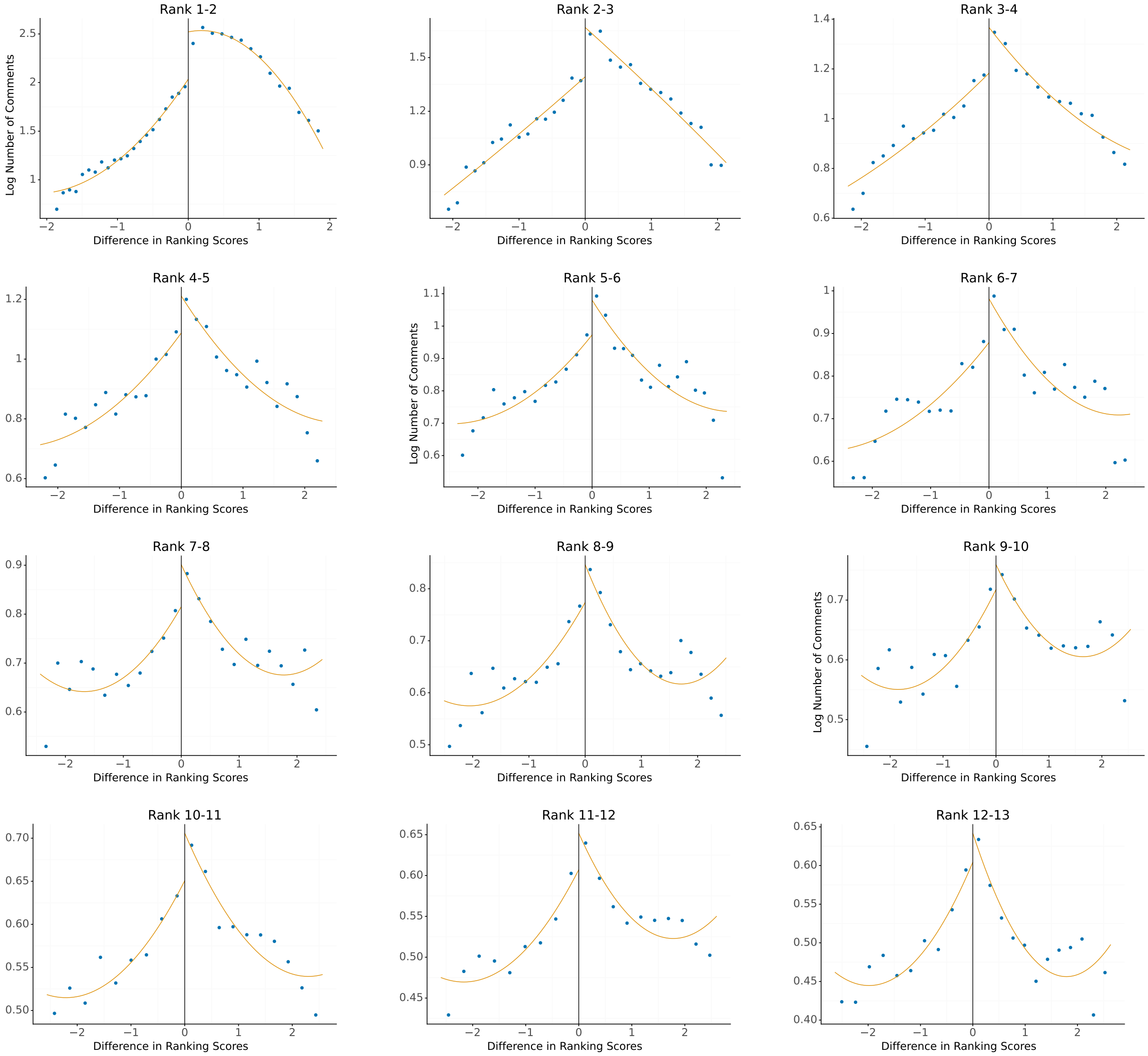
Note: Regression discontinuity first stage plots of the probability a post is ranked lower on the feed against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

Figure A.3: Regression Discontinuity Plots: First Stage



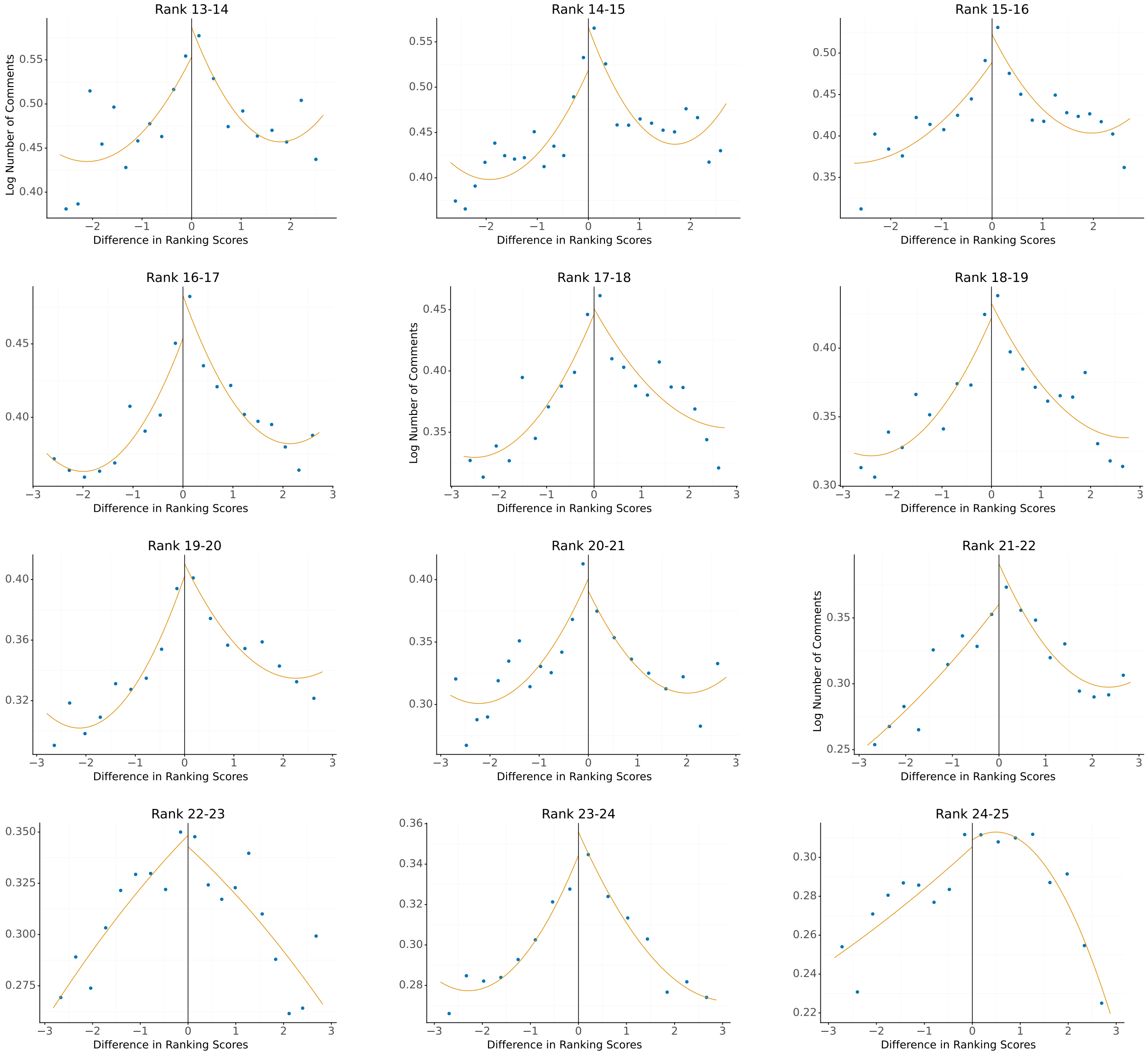
Note: Regression discontinuity first stage plots of the probability a post is ranked lower on the feed against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

Figure A.4: Regression Discontinuity Plots: Engagement



Note: Regression discontinuity outcome plots of the log number of comments received in the 20 minutes following a snapshot against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

Figure A.5: Regression Discontinuity Plots: Engagement



Note: Regression discontinuity outcome plots of the log number of comments received in the 20 minutes following a snapshot against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

A.2.2 Robustness of Regression Discontinuity

Regression Discontinuity with Two-Dimensional Score

Recall the running variable in the regression discontinuity analysis is a composition of two continuous scores, the difference in vote scores and the difference in post age. Figure A.6 plots the joint distribution of these two scores along the discontinuity frontier.

Balance of Covariates

Here, I show evidence that pre-treatment observable post features are continuous through the discontinuity. I show the full regression discontinuity plots for the top 12 positions on the feed for post vote score (Figure A.7), post age (Figure A.8), publisher slant (Figure A.9), and publisher credibility (Figure A.10). Estimates of the local average treatment effect on each of these covariates using local linear regression are displayed in Figure A.11.

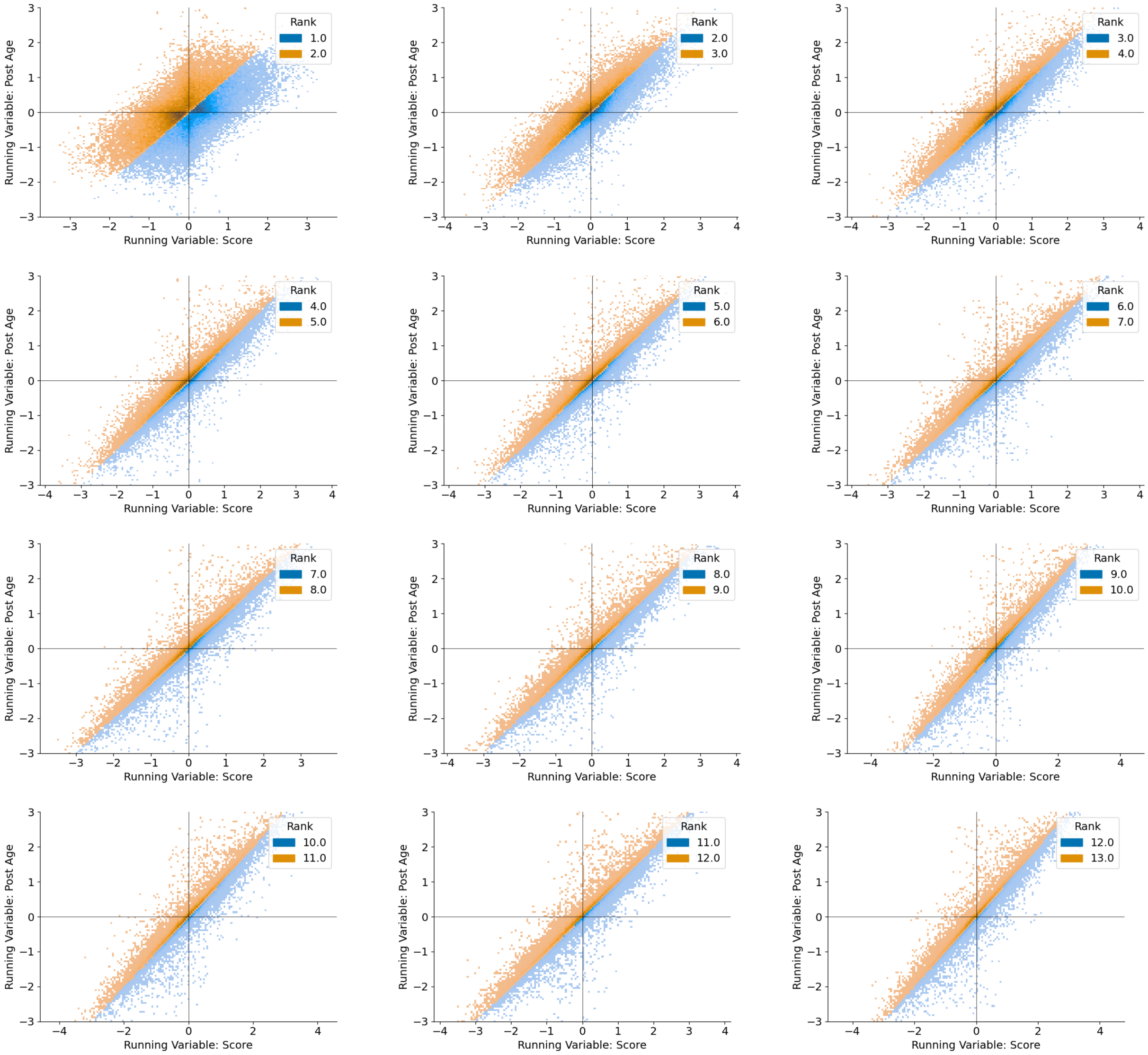
Robustness of Bandwidth, Donut, and Comment Window

Here, I show the position effect estimates are robust to researcher choices regarding the regression discontinuity bandwidth (Figure A.12), the donut of data excluded around the discontinuity (Figure A.13), and the window of comments included after a post snapshot (Figure A.14).

A.2.3 Recommender System Appendix

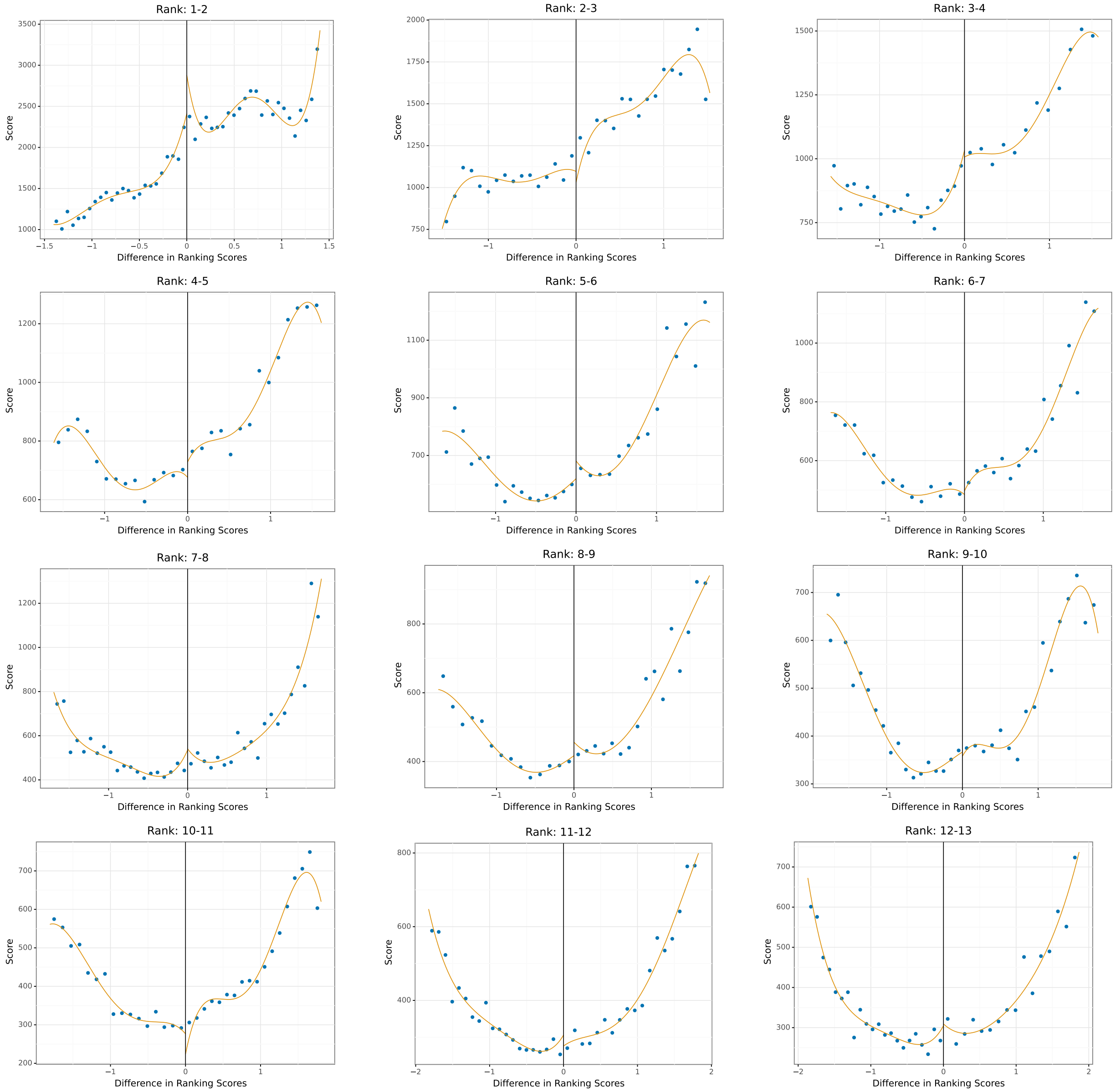
Table A.3 shows the projection of the first 3 principal components of the publisher features learned in the collaborative filtering model onto the vector of publisher ratings. These regressions demonstrate that the publisher ratings do explain some of the variation in the publisher features learned by the recommender system.

Figure A.6: Regression Discontinuity with Multiple Scores



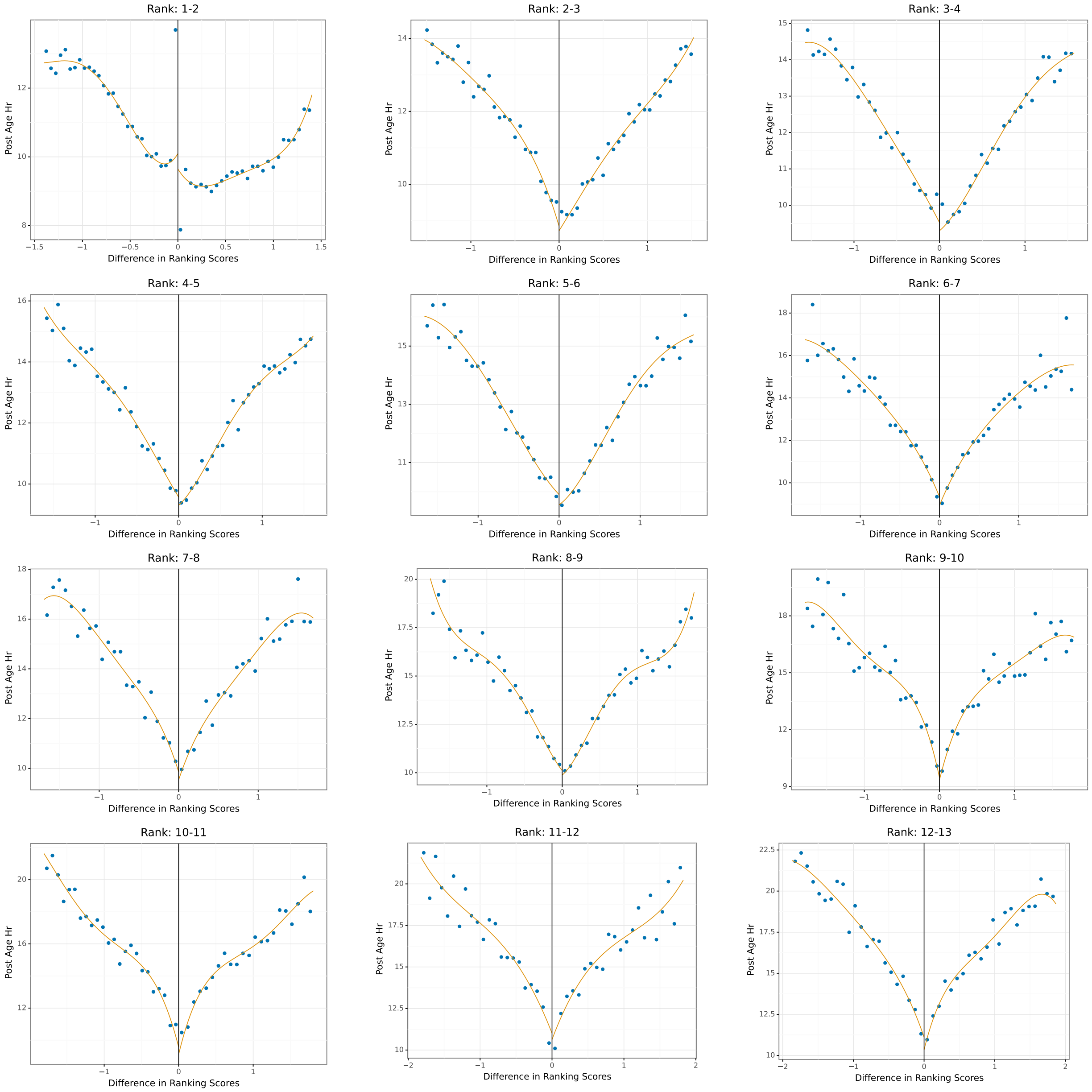
Note: This plot shows the regression discontinuity in two dimensions. The. x axis plots the difference in the normalized post vote scores and the y axis plots the difference in the normalized post ages. The discontinuity frontier corresponds to the 45 degree line. To make the charts easier to view, I only plot posts that are correctly classified by the running variable.

Figure A.7: Balance of Vote Score



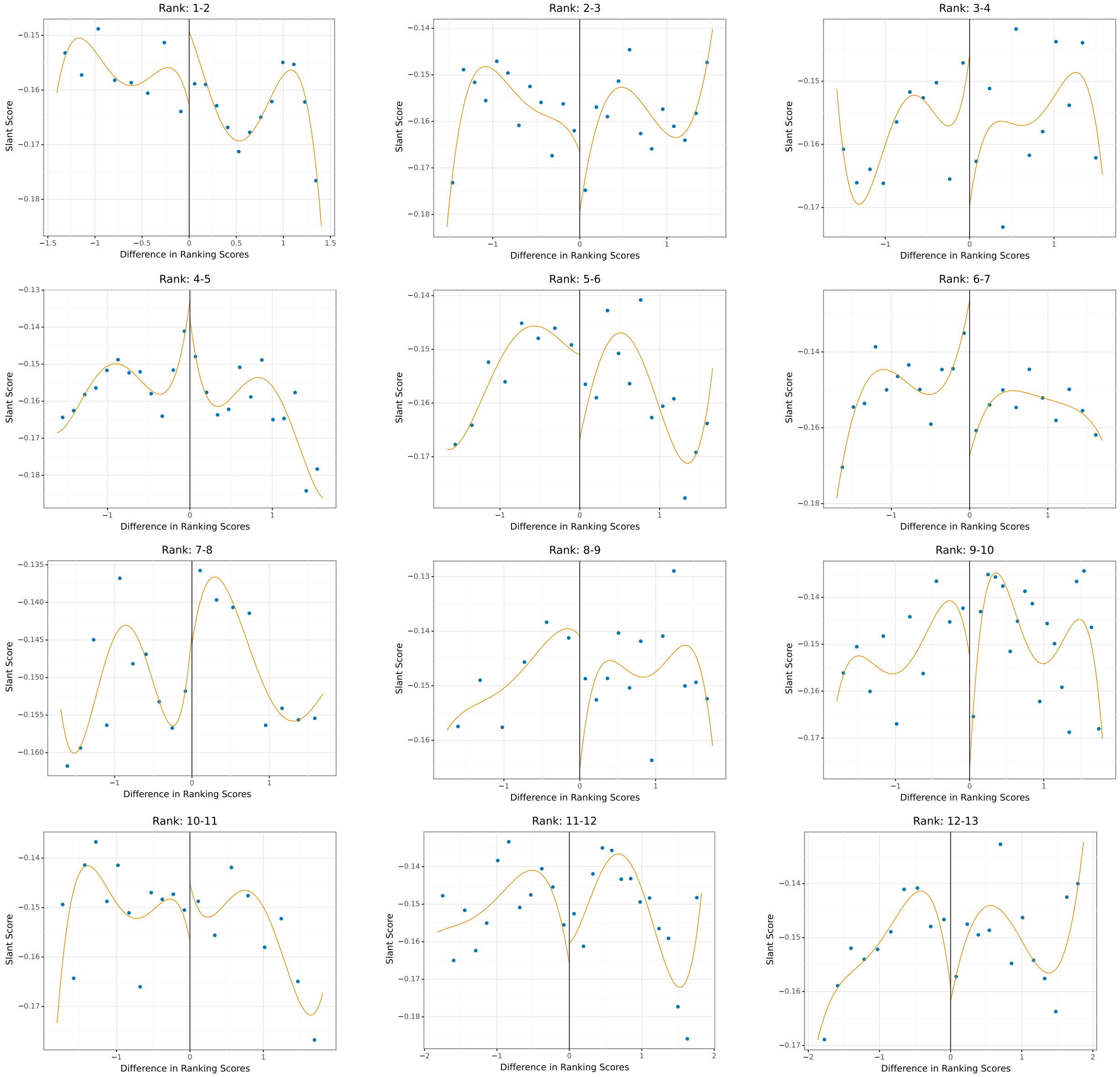
Note: This plot shows the binned means of post vote score against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure A.8: Balance of Post Age



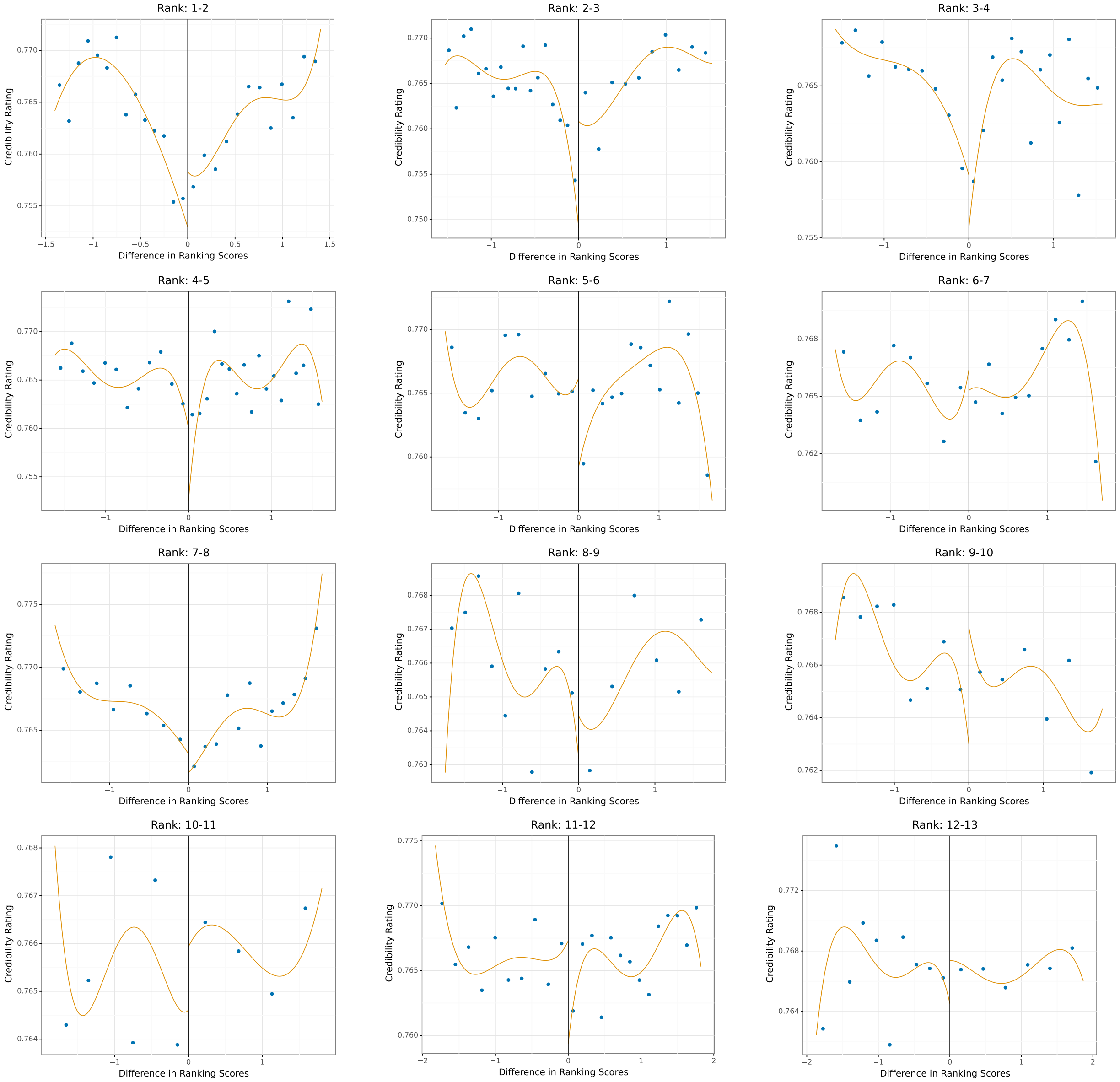
Note: This plot shows the binned means of post age against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure A.9: Balance of Slant Score



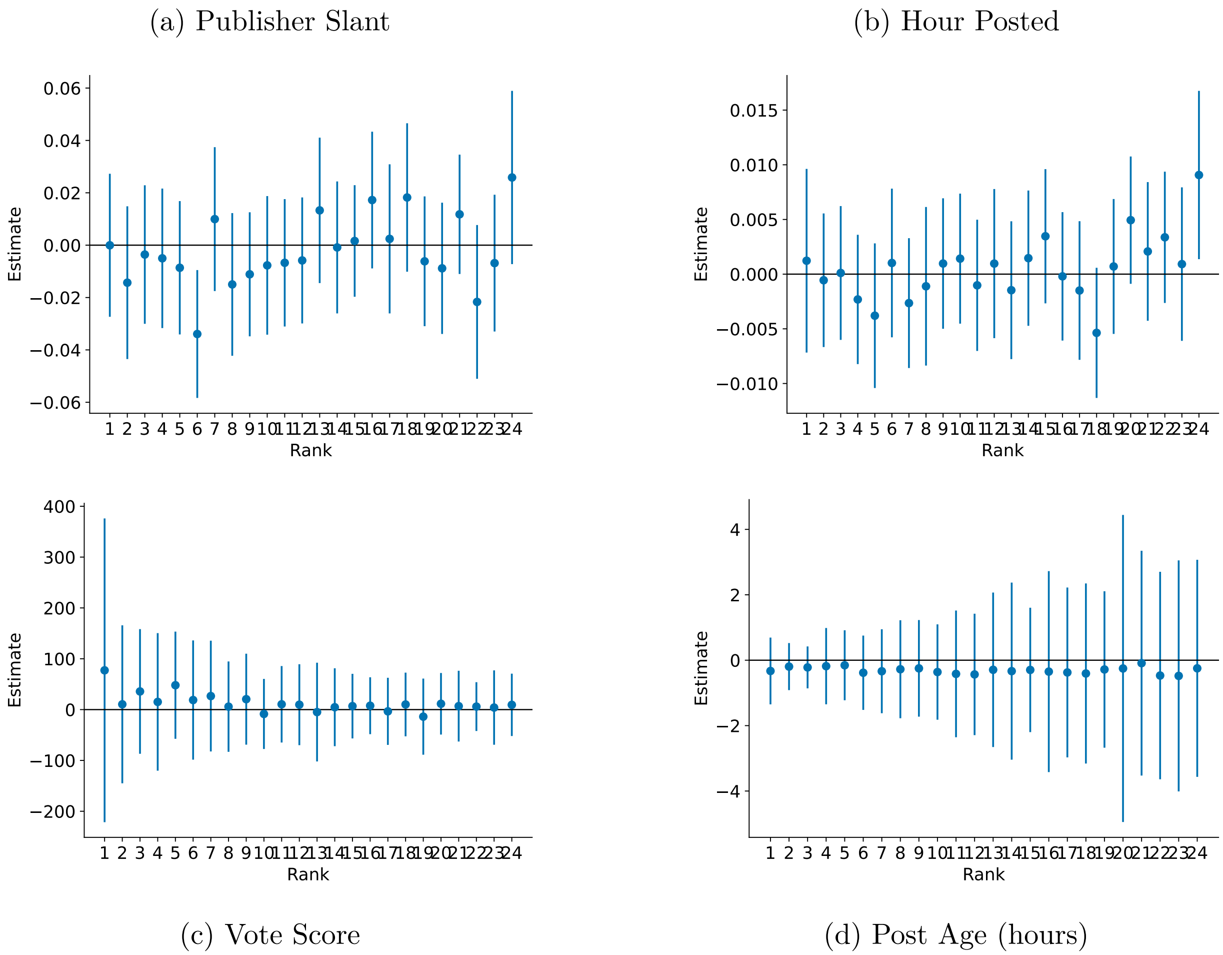
Note: This plot shows the binned means of publisher slant score against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure A.10: Balance of Credibility Rating



Note: This plot shows the binned means of publisher credibility rating against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure A.11: Regression Discontinuity Placebo Tests



(a) Publisher Slant

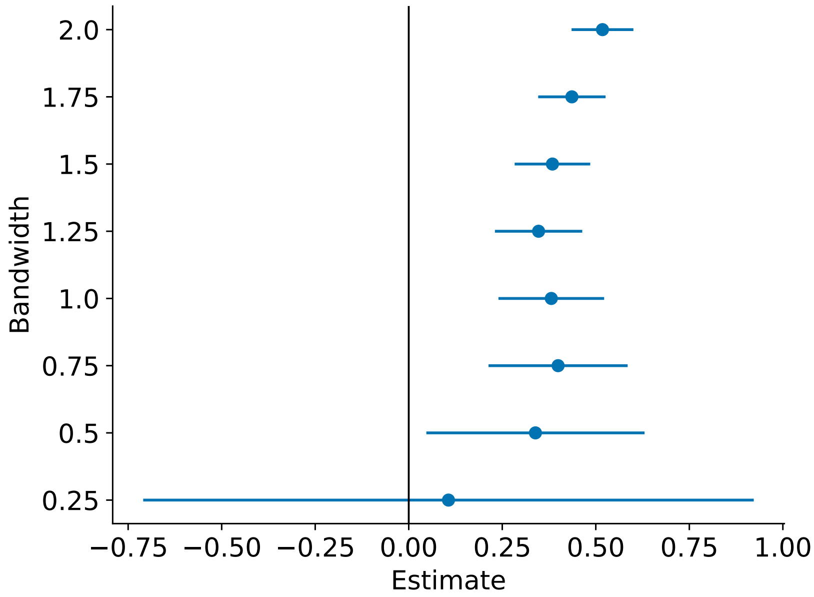
(b) Hour Posted

(c) Vote Score

(d) Post Age (hours)

Note: Placebo test for discontinuity of observable pre-treatment covariates. Each figure plots local average treatment effect estimates of moving from rank $r + 1$ to rank r using a local linear regression for publisher slant score, hour posted, vote score, and post age.

Figure A.12: Robustness of Position Effect Estimates to Bandwidth



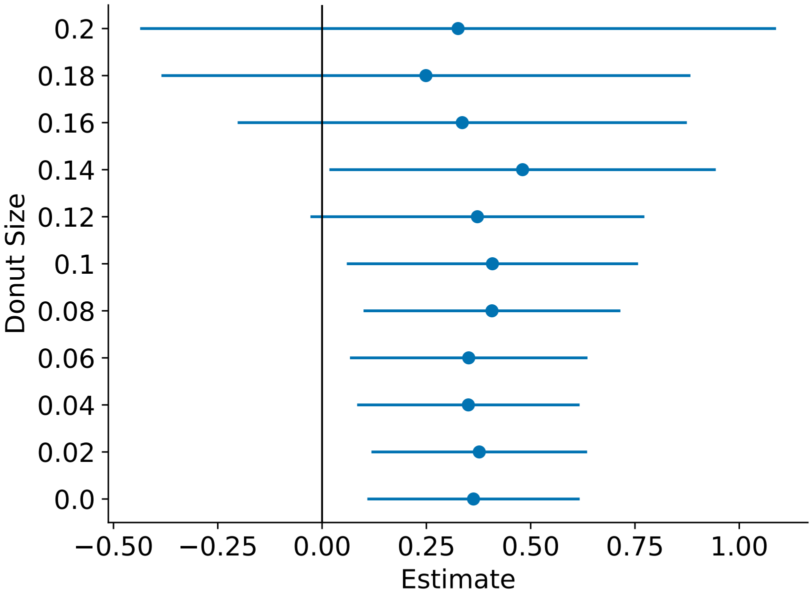
Note: This plot shows the robustness of the position effect estimate to bandwidth size. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log number of comments a post receives. Error bars represent 95% confidence intervals using robust bias-corrected standard errors.

Table A.3: Recommender System Publisher Factors

	(1)	(2)	(3)
Slant Score	-0.19***	-0.01	-0.07**
	(0.03)	(0.03)	(0.03)
Credibility Rating	-0.00	-0.00	-0.01
	(0.04)	(0.03)	(0.03)
Average Rank	0.00	-0.01	-0.02
	(0.02)	(0.01)	(0.02)
Quantity	0.05***	0.58***	0.17**
	(0.02)	(0.20)	(0.08)
Intercept	-0.07**	-0.04	0.20***
	(0.03)	(0.02)	(0.03)
Obs	1378	1378	1378
R^2	0.04	0.34	0.04

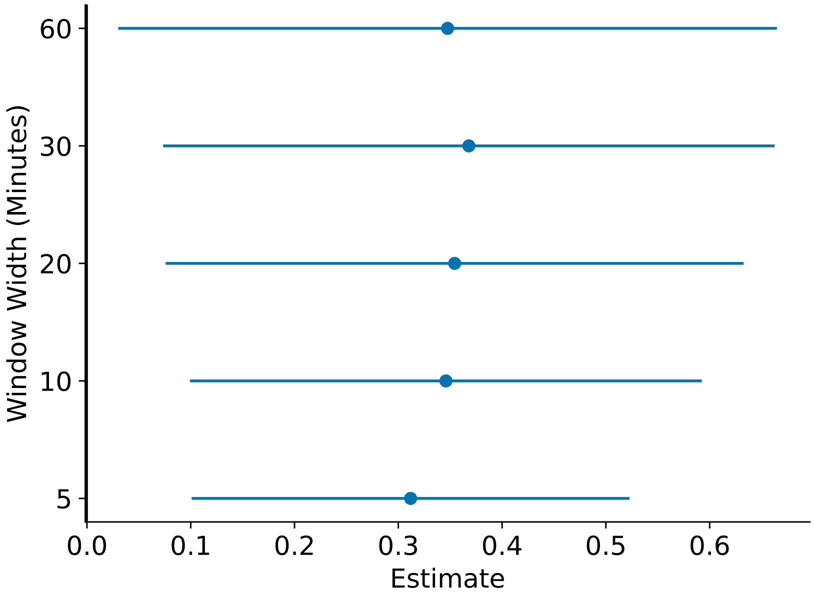
Note: This table shows estimates from a regression of the first 3 principal components of publisher features on publisher observables.

Figure A.13: Robustness of Position Effect Estimates to Donut Width



Note: This plot shows the robustness of the position effect estimate to donut size. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log number of comments a post receives. Error bars represent 95% confidence intervals using robust bias-corrected standard errors.

Figure A.14: Robustness of Position Effect Estimates to Comment Window

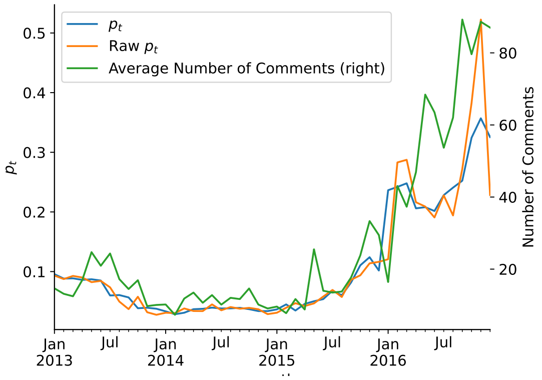


Note: This plot shows the robustness of the position effect estimate to window length. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log number of comments a post receives in the window following a snapshot. Error bars represent 95% confidence intervals using robust bias-corrected standard errors.

A.3 Choice Model Appendix

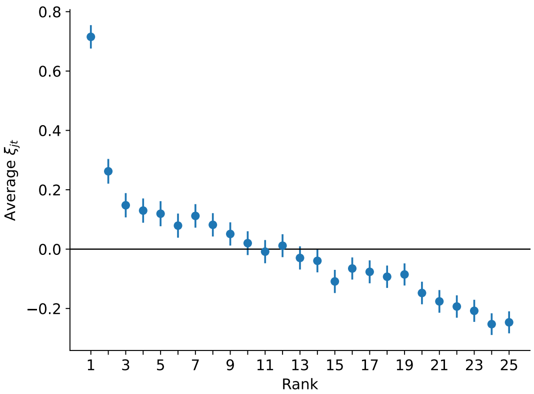
A.3.1 Additional Figures and Tables

Figure A.15: p_t and Average Top-Post Engagement



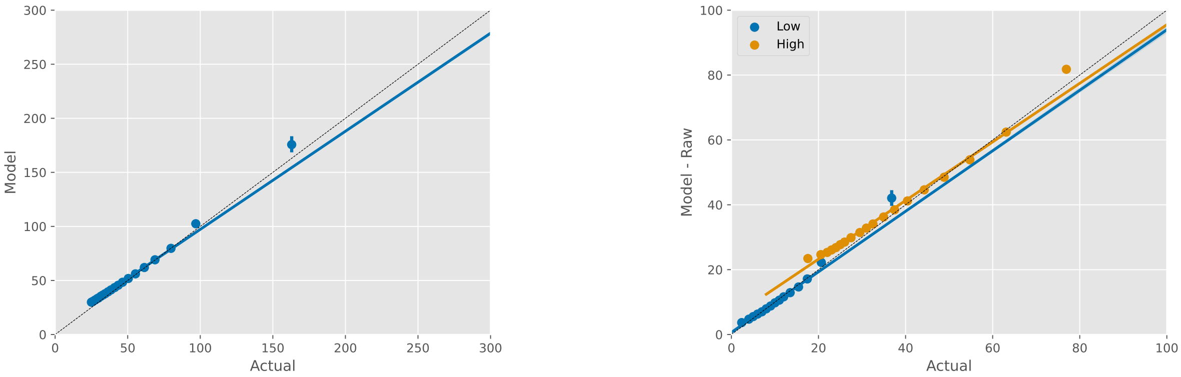
Note: This plot shows the time series of the raw and smoothed p_t estimates (left axis). The right axis shows the average number of comments the top post on the feed receives from the users in the sample.

Figure A.16: Average ξ_{jt} by Rank



Note: This plot shows the average ξ_{jt} value by post rank for the posts in the sample. Bars represent 95% confidence intervals.

Figure A.17: Summary of Model Fit



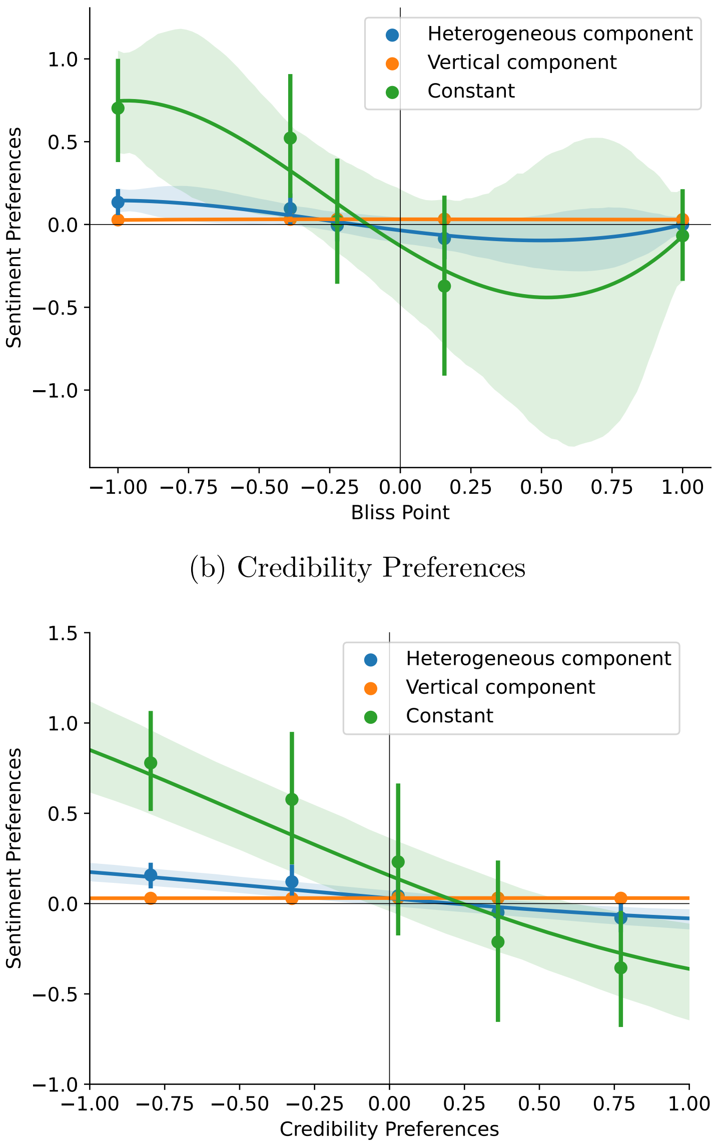
(a) Total Number of Comments

(b) Engagement by Credibility

Note: This plot shows additional summaries of the model fit. (a) The relationship between the actual number of comments a user posts against the model fitted number of comments the user submitted. (b) The same relationship, broken out by publisher credibility.

Figure A.18: Correlation of Sentiment Preferences with Comment Preferences

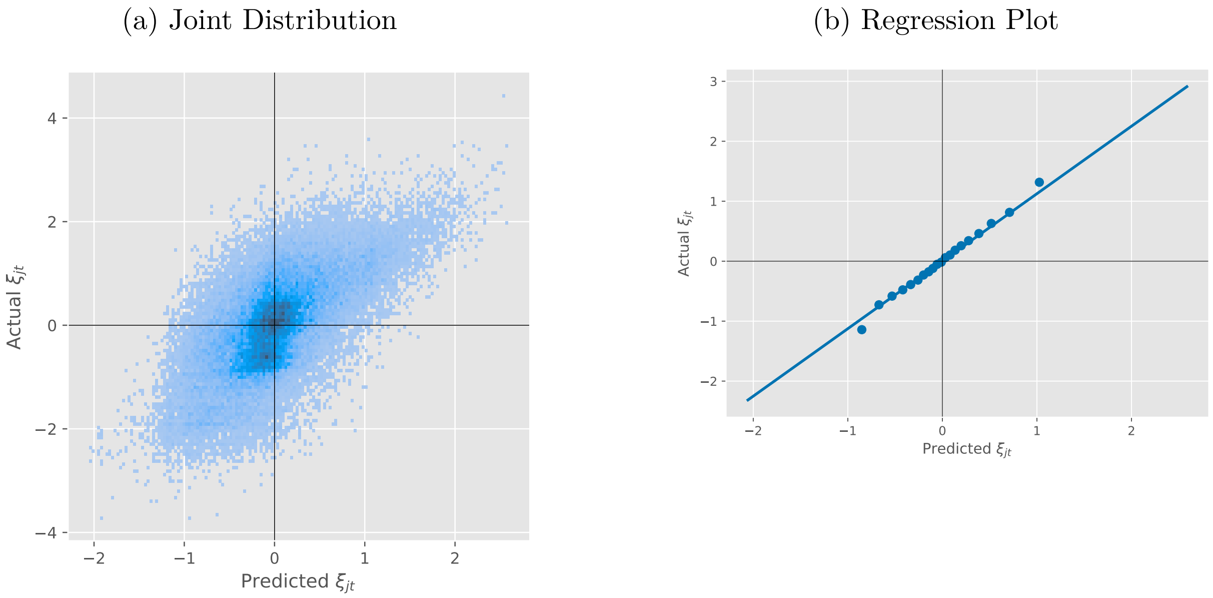
(a) Slant Bliss Point



(b) Credibility Preferences

Note: This figure plots binned mean sentiment preferences against (a) user bliss points and (b) credibility preferences. The heterogeneous component captures how the likelihood of a user to submit a negative comment changes in response to changes in the user-specific component of post comment utility. The vertical component captures how the likelihood of a user to submit a negative comment changes in response to a change in the latent commentability term (ξ_{jt}). Positive values mean the user is more likely to submit a negative comment on articles they are likely to comment on. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Figure A.19: Summary of ξ_{jt} Model

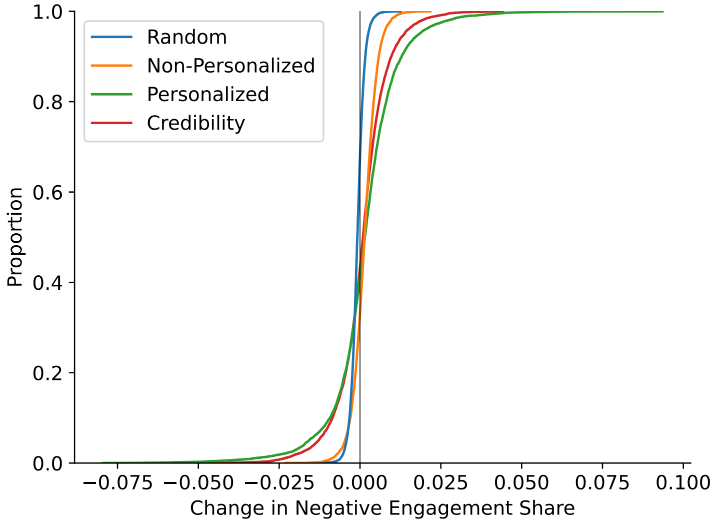

(a) Joint Distribution

(b) Regression Plot

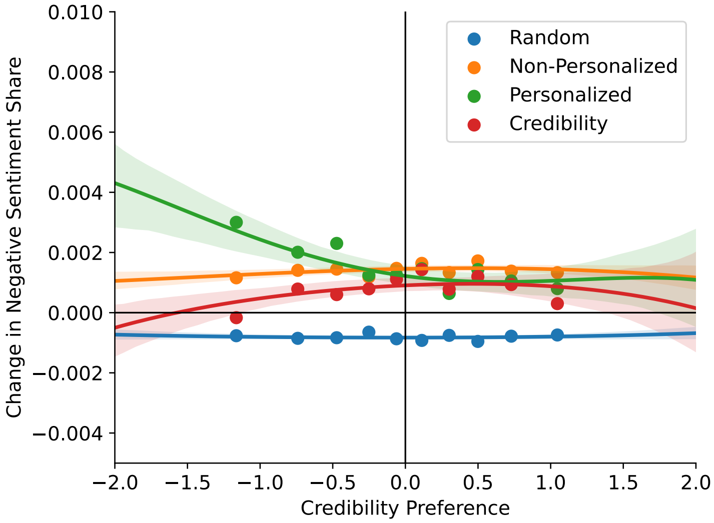
Note: This plot shows a summary of the random forest model used in the counterfactuals to estimate ξ_{jt} in each period. (a) shows the joint distribution between ξ_{jt} and $\hat{\xi}_{jt}$ and (b) shows binned means of this relationship along with a linear fit.

Figure A.20: Impact of Algorithm on Negative-Sentiment Share

(a) Distribution in Change in Negative-Sentiment Share



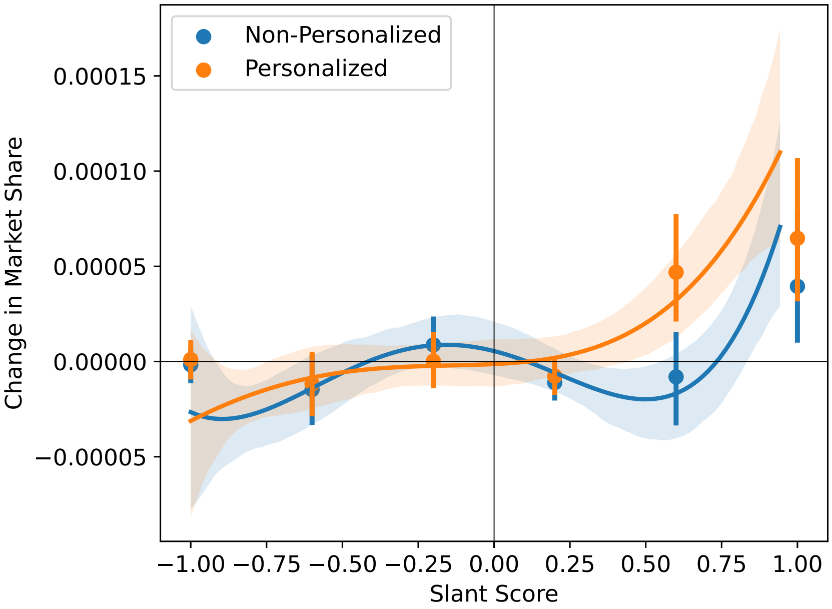
(b) Change in Negative-Sentiment Share by Credibility Preference



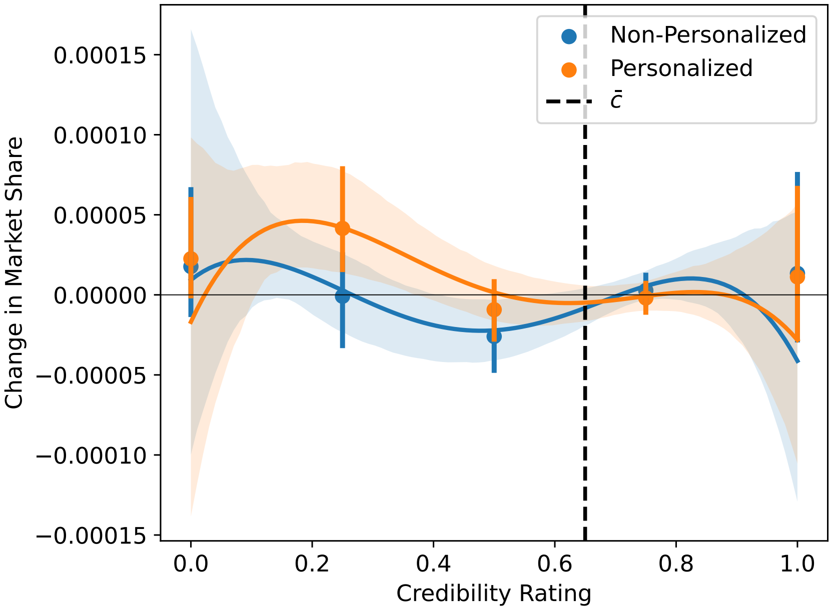
Note: (a) Plots the empirical CDF of the change in the share of users' comments that are negative sentiment under the counterfactual algorithms relative to the existing algorithm. (b) Plots the binned mean in users' change in negative sentiment score against their preferences for publisher credibility. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Figure A.21: Change in Publisher Market Shares

(a) Publisher Slant



(b) Publisher Credibility



Note: Figure A.21a plots the binned mean change in publisher market share by publisher slant and Figure A.21b plots the binned mean change in publisher market share by publisher credibility rating. In both figures, the regression lines are fourth-order polynomial fits. Confidence bands represent 95% confidence intervals.

A.3.2 Empirical Bayes Shrinkage

To shrink individual preference estimates towards the grand mean and adjust for the over-dispersion due to sampling error I use the following empirical Bayes procedure. I assume that the true individual preference parameters are drawn independently and identically distributed from a multivariate normal distribution

$$\beta_i \sim N(\mu, \Sigma)$$

and we observe noisy estimates of these parameters $\hat{\beta}_i = \beta_i + \nu_i$ where $\nu_i \sim N(0, \Sigma_i)$ is independent sampling error and Σ_i are estimated covariance matrices of preferences for each user. I form estimates of the grand-mean and covariance matrix using empirical analogs of the following expectations.[1]

$$\mu = E\left[\hat{\beta}_i\right]$$

$$\Sigma = E\left[\left(\hat{\beta}_i - \mu\right)\left(\hat{\beta}_i - \mu\right)'\right] - E\left[\Sigma_i\right]$$

and then form estimates of the posterior mean for each β_i as

$$E\left[\beta_i | \hat{\beta}_i, \Sigma_i, \mu, \Sigma\right] = \left(\Sigma^{-1} + \Sigma_i^{-1}\right)^{-1}\left(\Sigma^{-1}\mu + \Sigma_r^{-1}\hat{\beta}_i\right).$$

This shrinks each estimated preference parameter towards the grand mean and corrects for the over-dispersion created by sampling error.

A.3.3 Estimating the Share of Users Accessing the Platform

Little's law shows that in a stationary system, the average number of users on the platform can be expressed as

$$L_t = \lambda_t W \tag{A.1}$$

where L_t is the average number of users on the platform at any point during period t, λ_t is the arrival rate of customers during period t, and W is the average session length [Little, 1961]. I assume $W = 10.82$ given that the average session length on Reddit in 2016 lasted 10 minutes and 49 seconds.[2] I assume that the number of users $A_t^0 = L_t$: the number of users online at the start of each period is equal to the average over the period. I can re-arrange equation A.1 to show that $A_t = \frac{l}{W} A_t^0$ which says the total number of users to visit the platform during period t (A_t) equals the length of the period in minutes (l) divided by the

[1] When estimating the grand mean, I use inverse variance weights to improve precision of the estimated mean.

[2] https://web.archive.org/web/20161203082123/https://www.similarweb.com/website/reddit.com/

session length (W) multiplied by the number of users online at any given time. To calibrate the number of active community members in a subreddit, I use two snapshots of the politics community's usage statistics from 2015 and 2016 to calculate the average number of unique users per day.[3] When combined with the number of subscribers a community has, I can estimate the share of subscribers that are active in a given day which averages 0.071 over the months covered in the two snapshots. I then calculate $N_t = 0.071 \times S_t$ where S_t is the number of subscribers the community has at period t. Robustness to the scaling factor is shown in Appendix Section A.3.4. Finally, I smooth estimates of p_t by taking the fitted values of the following regression model

$$\frac{A_t}{N_t} = \gamma_0 + \gamma_{quarter} + \gamma_{day} + \eta_t \tag{A.2}$$

where $\gamma_{quarter}$ and γ_{day} are quarter and day of week fixed effects, respectively.

A.3.4 Choice Model and Counterfactual Robustness

Robustness to Scaling $p(\cdot)$

First, I show that the counterfactual results are robust to scaling the exposure probability $p(\cdot)$ in the choice model. This shows the results are robust to the decision to scale the number of active users in Section A.3.3 and to the assumption that all users are exposed to the first post in the feed ($p_1 = 1$) as both simply multiply either p_t or p_r by a constant, so showing $p(\cdot)$ is robust to being multiplied by a constant demonstrates robustness to both.

[3] https://web.archive.org/web/20160905095430/https://www.reddit.com/r/politics/about/traffic
https://web.archive.org/web/20150513102644/http://www.reddit.com/r/politics/about/traffic/

Table A.4: Counterfactual Engagement Summaries Robustness: $p'(\cdot) = 0.5p(\cdot)$

	Engagement	Diversity	Max Partition Share	Credibility	Negative Engagement Share
Intercept	53.537	1.519	0.290	0.791	0.512
	(0.376)	(0.000)	(0.000)	(0.000)	(0.002)
Random	-6.144	0.005	-0.001	-0.000	-0.001
	(0.037)	(0.000)	(0.000)	(0.000)	(0.000)
Non-personalized	11.211	0.003	-0.005	0.010	0.002
	(0.057)	(0.000)	(0.000)	(0.000)	(0.000)
Personalized	12.270	-0.017	0.021	0.000	0.002
	(0.066)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	33340	33340	33340	33340	33340
R-Squared	0.043	0.048	0.074	0.006	0.000

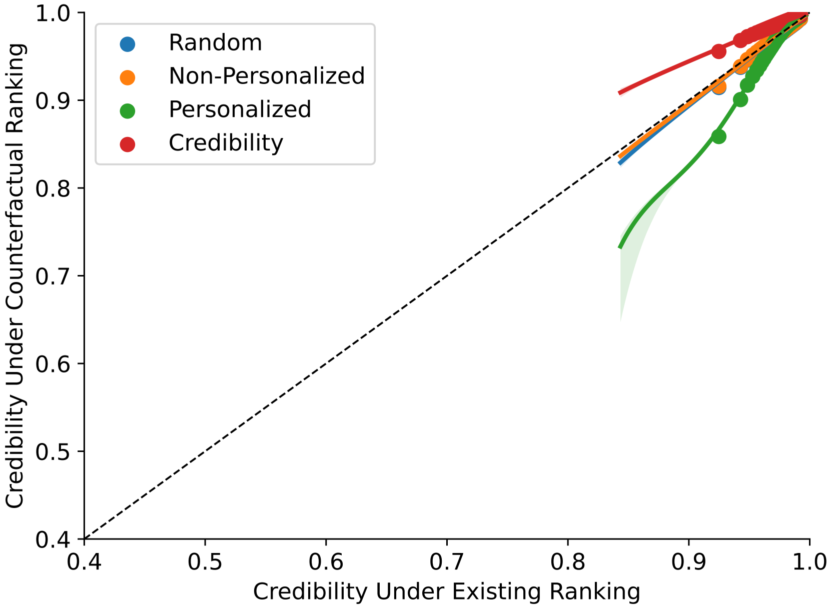
Note: This table reports estimates of a panel regression of each counterfactual outcome on counterfactual algorithm dummy variables in the robustness exercise where $p(\cdot)$ is multiplied by a factor of 0.5. The intercept is the average quantity under the existing algorithm. Standard errors are clustered at the user level.

Robustness to Choice of $\underline{c}$

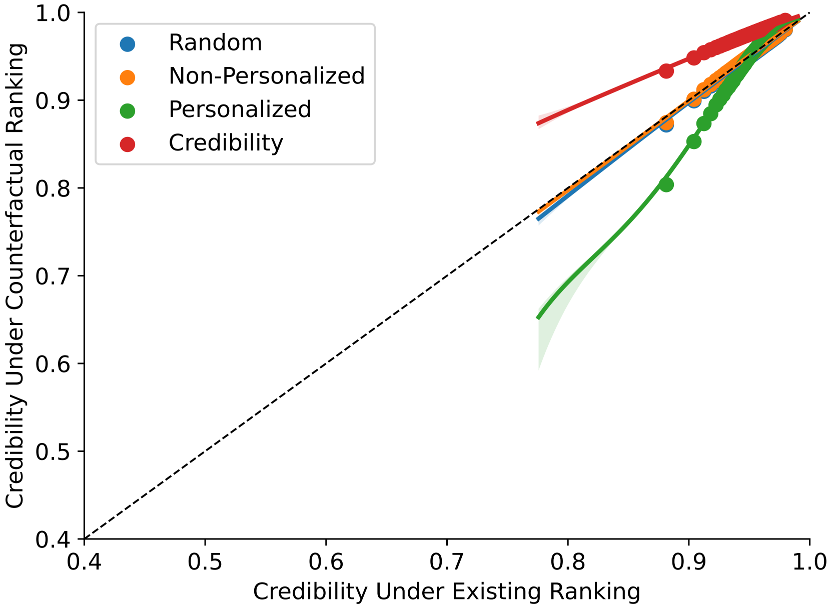
I replicate the quality analysis for various choices of $\underline{c}$ and find the results are qualitatively similar (Figure A.22). For values of $\underline{c}$ as low as 0.4, I find the personalized engagement maximizing exacerbates differences in users along the credibility dimension. As the threshold for credibility is lowered, by definition the share of high quality engagement rises as the threshold does not impact the counterfactuals directly, only how the counterfactuals are evaluated.

Figure A.22: High-Credibility Share by Baseline Credibility: Robustness

(a) $\underline{c} = 0.4$

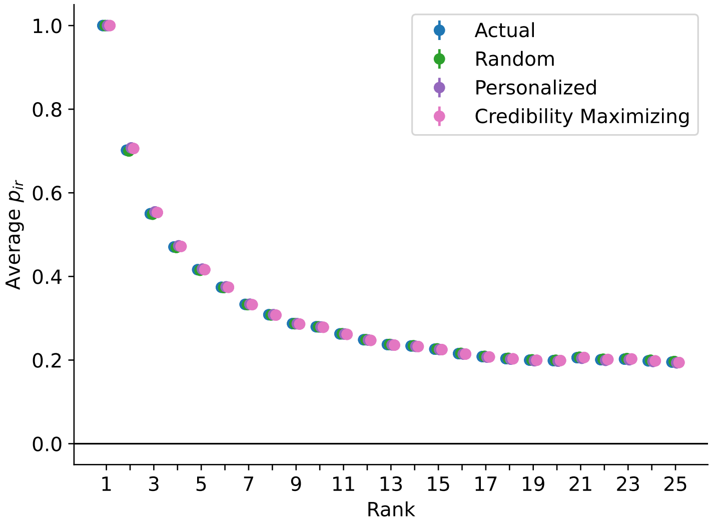


(b) $\underline{c} = 0.5$



Note: This figure plots binned mean credibility shares under the counterfactual algorithm against credibility shares under the existing algorithm for various thresholds of high-quality publishers. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Figure A.23: Reoptimized Exposure Probabilities



Note: Average exposure probabilities ($E\left[P\left(\bar{u}_{ir_r} > c_{ir_j} + \eta_{ijt}\right)\right]$) in the endogenous search model by position and counterfactual algorithm.

Robustness to Endogenous Search

A.4 A Recommender System Approach to Personalization

In this section, I study the types of content that are promoted when personalizing the ranking algorithm to maximizing engagement using a reduced form approach. To explore this, I train a collaborative-filtering based recommender system using the matrix of user-level comment counts by publishers. The recommender system then recommends publishers on which a user is most likely to comment in a period. I validate this recommender system by estimating heterogeneous treatment effects when the regression discontinuity experiments align with the recommender system's predictions. I find that the recommender system effectively predicts treatment effects, a result that suggests the model has learned important aspects of user preferences. I then study the types of content that gets promoted under this simple recommender system to understand the extent to which personalized engagement maximization impacts individual news diets.

The primary purpose of this section is to provide reduced-form evidence that personalizing content to maximize engagement promotes low-credibility content to a subset of users and lowers the diversity of publishers that are promoted. This approach has two advantages over the discrete choice model and counterfactual analysis I study in Section 1.4 and Section 1.5. First, this model is trained using comment decisions from over 500,000 users and is evaluated on comment decisions of over 180,000 users. This is a much larger sample than that used in the choice model approach, as I can use comment decisions on articles during periods not captured in Wayback Machine snapshots during the training process. I find consistent results across both approaches, which gives confidence that the findings of the choice model approach can be generalized to a broader set of users. Second, I can evaluate this simple approach by predicting treatment effects to give confidence that the model has learned important aspects of preferences.

A.4.1 Training and Validating the Recommender System

To train the recommender system, I split user-level comment data into training and test sets. The test set consists of comments on articles that appear in Wayback Machine snapshots and the training set consists of comments on articles that do not appear in Wayback Machine snapshots. The test set is used to evaluate the recommendations through heterogeneous treatment effects. I focus this analysis on the politics community because of its importance to managers, policy makers, and users. In the training set, I generate a matrix of user comment counts by publisher domain, where each row represents a user and each column a publisher. I use this matrix to train a collaborative filtering model for implicit data, following Hu et al. [2008]. This simple model assumes that user preferences for a publisher can be represented by the dot product of low-rank vectors of latent user and publisher features. Appendix A.2.3 shows the publisher embeddings learned by the model are correlated with observable features. More specifically, publisher popularity and slant are the observable publisher features most correlated with the latent embeddings. Given the publisher and user features, the recommender system then recommends publishers that a user is more likely to prefer.

To evaluate the recommender system, I estimate heterogeneous treatment effects comparing periods when the model predicts a user's preferred publisher was promoted to the top of the feed in the regression discontinuity experiments relative to when the non-preferred publisher was promoted. For each period and user, I determine if the preferred post of a user is promoted, the preferred post is demoted, or the user is indifferent. A user is indifferent in a period if the two publishers are within 1 percentile of one another in the model's recommendations for that user. The preferred publisher is promoted for a user in a period if the publisher of the first post is at least 1 percentile higher than the publisher of the second post in the model's recommendations for the user. Likewise, the preferred publisher is demoted if the publisher of the first post is at least 1 percentile lower than the publisher of the second post. I then sum the total number of comments for each post and period across users based on whether the user-period is classified as the preferred post being promoted, demoted, or indifferent. Finally, I estimate the regression discontinuity heterogeneous treatment effects through local linear regression. Given the reduced power in identifying heterogeneous treatment effects, I inflate the bandwidth used in Section 1.3.3 by a factor of two and use cluster robust standard errors rather than standard errors that are robust to misspecification of the conditional expectation function.[4]

Heterogeneous treatment effect estimates are shown in Table A.5 and suggest that the recommender system effectively predicts treatment effects for the top position in the feed. The treatment effect is substantially – 13 percentage points – larger when a user's preferred publisher is promoted versus when the user's preferred publisher is demoted. That the recommender system is able to predict treatment effects confirms that the recommender system has learned important aspects of user preferences.

A.4.2 Recommender System Results

I now turn to summarizing the properties of the recommender system to understand the types of content promoted when personalizing rankings and to motivate the choice model presented in Section 1.4. For each user-period, I determine the most preferred publisher according to

[4] This assumes that the conditional expectation function is linear within the bandwidth and does not account for misspecification.

Table A.5: Validating the Recommender System

	Preferred Promoted	Preferred Demoted	Indifferent
D	0.80	0.67	0.62
	(0.04)	(0.04)	(0.04)
Δs_{jt}	0.62	0.49	0.72
	(0.12)	(0.12)	(0.13)
$\Delta s_{jt} \times D$	-0.97	-0.57	-1.44
	(0.25)	(0.25)	(0.28)
Intercept	2.14	2.14	1.95
	(0.06)	(0.06)	(0.07)
Obs	3538	3538	3538
R^2	0.09	0.08	0.04

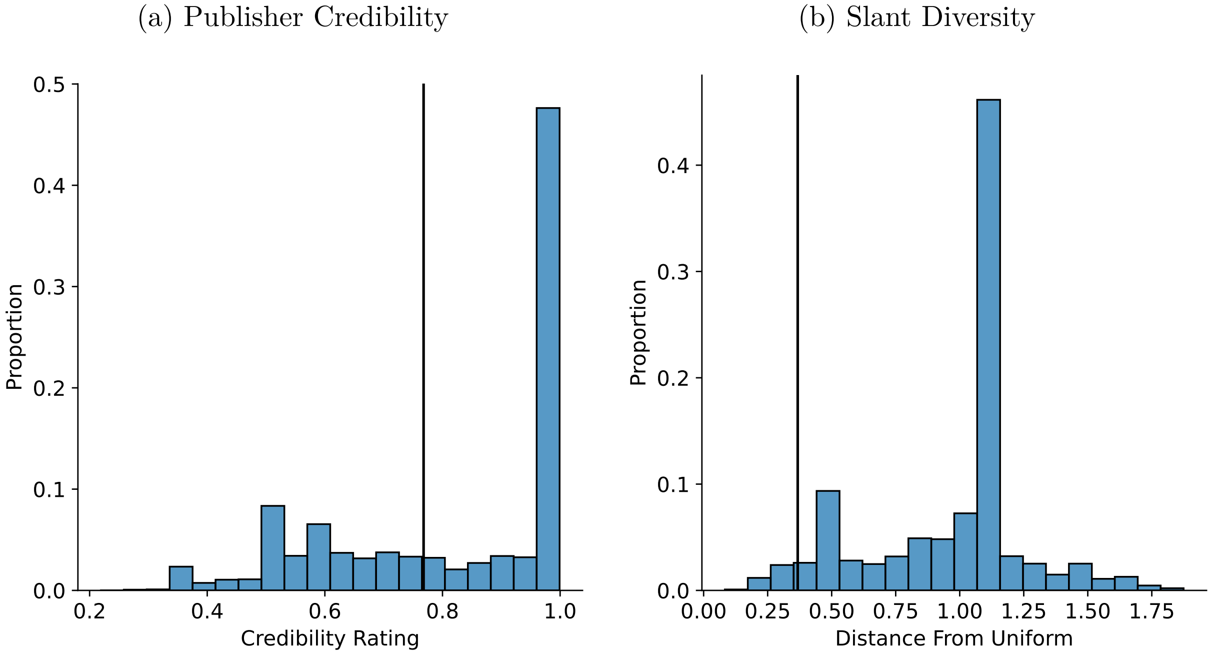
Note: This table shows regression discontinuity heterogeneous treatment effect estimates using a local linear regression. Each column presents estimates of the local linear regression of the outcome (log of one plus the number of comments of each type) on an intercept, treatment indicator, running variable, and running variable interacted with the treatment indicator. The first row contains the coefficient on treatment, which is the local average treatment effect of being promoted to the first position on the feed from the second position. The treatment effect is estimated separately depending on whether a user's preferred publisher is promoted (first column), demoted (second column), or the user is indifferent between the publishers (third column) in the given period.

the recommender system and calculate the share of promoted publishers that are classified as highly credible. I also calculate the primary measure of slant diversity, which is the first Wasserstein distance between the share of publishers promoted in each slant partition and the uniform distribution. The distributions of these summaries are shown in Figure A.24 alongside the quantity under the existing ranking. The distributions indicate that the majority of users experience improved news diet quality in terms of publisher credibility, though an important minority of users experience a material deterioration in the quality of their news diets. In terms of diversity, a large majority of users are recommended a less diverse set of publishers.

While these results suggest that optimizing for engagement using personalized rankings has a heterogeneous impact on the credibility of publishers that are promoted and a near-uniform decrease in slant diversity, this approach has important limitations. First, the recommender system is trained on observational data without accounting for endogenous post rank [Chaney et al., 2018]. The simple collaborative filtering model trained here also differs substantially from the more advanced models – which often employ deep learning – used in practice (see Zhang et al. [2019] for an overview of current deep learning based approaches to recommendation systems). In addition, this approach does not allow for within-publisher article heterogeneity, wherein certain articles are likely to garner more attention irrespective of the publisher. Finally, this approach is limited to analyzing the type of content promoted rather than modeling the content users eventually engage with under counterfactual rankings. Because it allows me to quantify the counterfactual ranking algorithms' impact on engagement – an outcome that serves as a closer proxy to advertising revenue – modeling engagement is critical to understanding the implications for the platform. The choice model and counterfactual analysis presented in Section 1.4 and Section 1.5 address these limitations directly.

Despite these limitations, that this model can accurately predict treatment effects indicates the model has learned useful information about user preferences. Moreover, this model can be estimated using data from a larger set of users since engagement on posts not included in the Wayback Machine snapshots can be included in training. This allows the recommender system approach to encompass over 180,000 users while the choice model is estimated on a smaller set of highly active users. As I will argue, the results from both analyses are similar and give confidence that the results generalize to a broader set of users.

Figure A.24: Summary of Promoted Publishers in Recommender System Approach


(a) Publisher Credibility

(b) Slant Diversity

Note: This figure summarizes the user-level distribution of promoted publishers in the recommender system approach to personalization. In each user-period, I find the publisher out of the top 25 posts that the recommender system would promote first. These figures plot the user-level distribution of the high-credibility publisher share and the first Wasserstein distance between the share of promoted publishers from each slant partition and the uniform distribution. The distance is zero when a user is equally likely to be promoted a publisher from each partition of publisher political slant. Higher values of the Wasserstein distance indicate the user is being promoted a less diverse set of publishers. The maximum distance is 2, which would only occurs when a user is promoted entirely publishers from either the extreme left or extreme right publisher partitions.

Appendix B

Appendix: News Feeds and User Engagement: Evidence from the Reddit News Tab

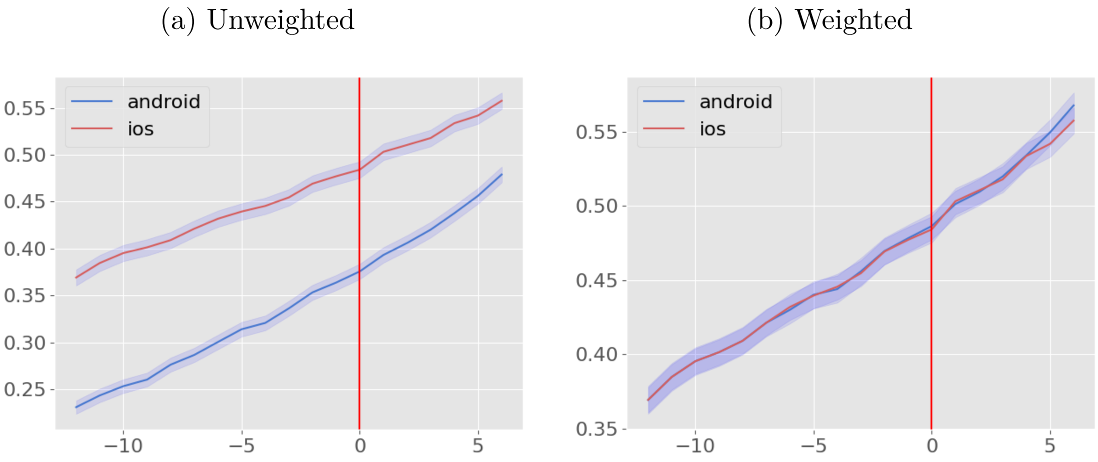
B.1 Matching Procedure

Our analysis of the impact of the News tab relies on a common trends assumption to have a causal interpretation. Non-weighted engagement trends, i.e. those estimating Equation 2.2 without the Coarsened Exact Matching (CEM) procedure are largely consistent with this hypothesis with one notable exception where the probability of posting on any non-news community shows a clear pre-trend (Figure B.11 and Figure B.12). We use the CEM procedure outlined in Iacus et al. [2012] to mitigate this imbalance.

There are 12 pre-treatment periods in our analysis and we match on the first 6 of these pre-treatment periods, holding out the following 6 to provide further evidence our common trends assumption is plausible. We define $X_i = (x_{i,-12}, \ldots, x_{i,-7})'$ as a vector of dummy variables $x_{i,t}$ equal to one if user i made any posts (on any community, not just news) in period t. We then generate weights from CEM following Iacus et al. [2012]. Given the relatively low-dimensionality of this matching exercise we are able to find exact matches and this is equivalent to subclassification. We then use these weights in our difference-in-differences framework, which requires the assumption of common trends conditional on the covariates used in matching. In other words, we assume that iOS engagement trends would have followed Android engagement trends absent the introduction of the News tab, conditional on X_i.

To demonstrate the effectiveness of the matching procedure, we can plot the unweighted and weighted probability of posting any post on Reddit over time by treatment group (Figure

Figure B.1: Probability of Any Post by Treatment Group


(a) Unweighted

(b) Weighted

B.1). While difficult to see in the raw time series figure below, there is clear evidence of differential pre-trends between iOS and Android users (Figure B.11). Re-weighting eliminates this imbalance and it does so even in the periods before treatment that are not used in weighting.

B.2 Data on Domain Political Slant

We obtain political slant by domain from the data provided alongside Robertson et al. [2018]. To generate this score, Robertson et al. [2018] collect recent tweets containing links from known Democrats and Republicans. The slant measure is then calculated as the difference in the probability of a sharing a domain conditioned on being republican who has shared at least one domain less the same conditional probability for democrats, normalized to be between -1 and 1. Formally, the slant measure is define as

$$\text{bias-score}(i) = \frac{\frac{r_i}{\sum_{j \in I} r_j} - \frac{d_i}{\sum_{j \in I} d_j}}{\frac{r_i}{\sum_{j \in I} r_j} + \frac{d_i}{\sum_{j \in I} d_j}}$$

where r_j (d_j) is the number of unique Republicans (Democrats) who shared domain j and I is the set of all domains. This measure is equal to 0 if shared by equal shares of Republicans and Democrats and equal to 1 (-1) if it was shared only by Republicans (Democrats). Robertson et al. [2018] demonstrate their measure of domain slant agrees with several existing measures of publisher slant [Bakshy et al., 2015b, Budak et al., 2016].

B.3 Measures of Information Diversity

Our measures of engagement diversity first partitions engagement into K bins based on either the category of news community or the political slant of the publisher within hard-news communities as calculated in Robertson et al. [2018]. We then operationalize the diversity of engagement following Holtz et al. [2020], measuring individual-level engagement diversity using Shannon entropy [Shannon, 1948]. The Shannon entropy of user i's engagement is defined as

$$id_i = -\sum_{k=1}^{K} s_{ki} \ln(s_{ki})$$

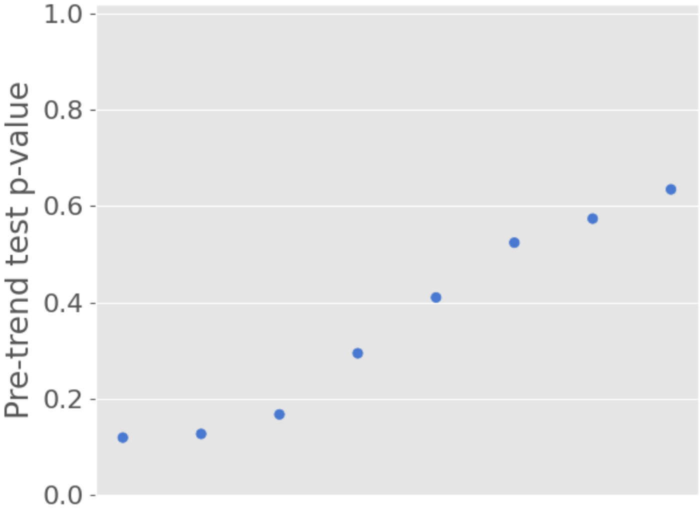
where s_{ki} is the share of user i's posts on a thread based on an article from publishers in partition $k \in \{1, \ldots, K\}$. If $s_{ki} = 0$, the partition's contribution to the Shannon entropy is zero which implies users who do not engage with any posts have $id_i = 0$ [Holtz et al., 2020].

We also operationalize engagement diversity using the Herfindahl–Hirschman Index (HHI) [Rhoades, 1993]. This measure is defined as the sume of squared engagement shares:

$$\text{HHI}_i = \sum_{k=1}^{K} s_{ki}^2.$$

When an individual has no consumption in a period, we define the HHI to be equal to one which is equivalent to engaging entirely with content from a single category.

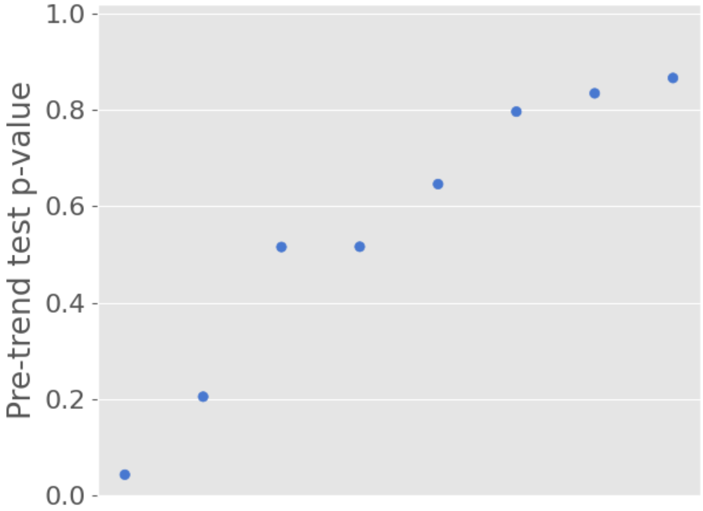
Figure B.2: Common Pre-Trends p-Values for Indicator of Any Post



Note: Plot of p-values of test of common pre-trends after CEM matching. The test is the joint test that all pre-period treatment effect coefficients in Equation 2.2 are equal to 0.

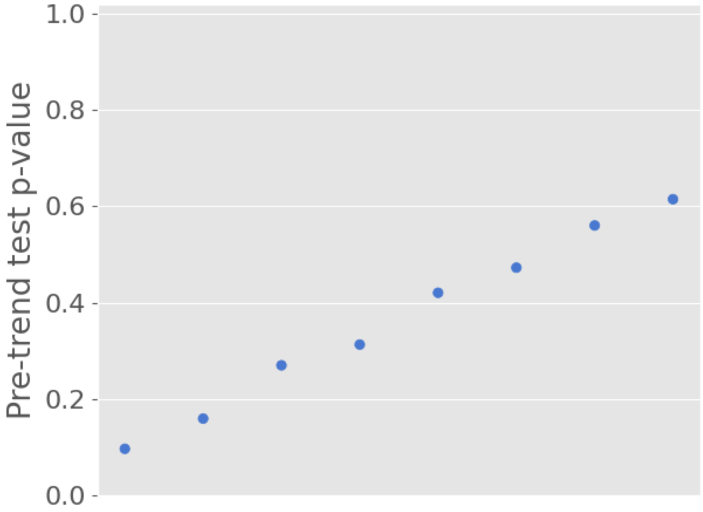
B.4 Testing for Common Pre-Trends

Figure B.3: Common Pre-Trends p-Values for Indicator of Low Quality Post



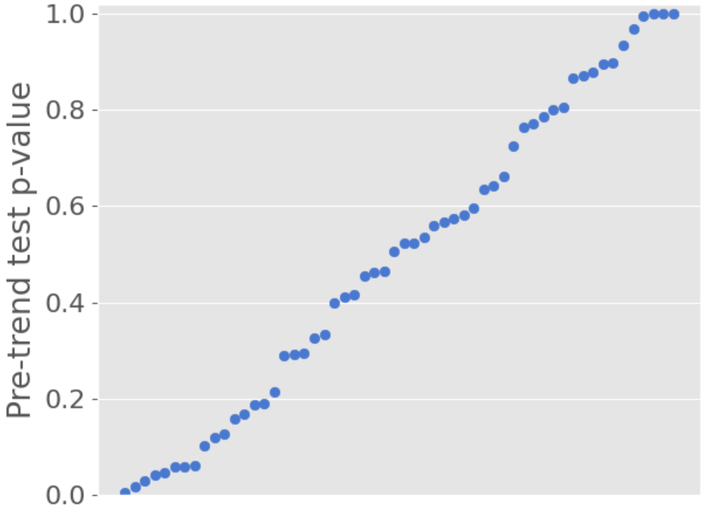
Note: Plot of p-values of test of common pre-trends after CEM matching. The test is the joint test that all pre-period treatment effect coefficients in Equation 2.2 are equal to 0.

Figure B.4: Common Pre-Trends p-Values for Indicator of High Quality Post



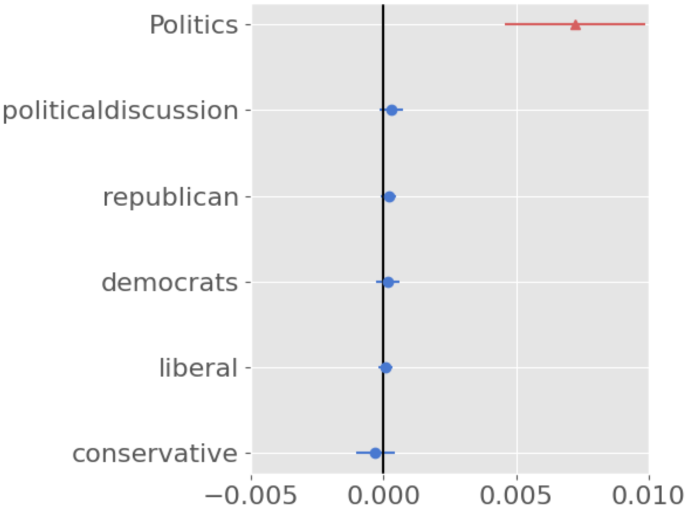
Note: Plot of p-values of test of common pre-trends after CEM matching. The test is the joint test that all pre-period treatment effect coefficients in Equation 2.2 are equal to 0.

Figure B.5: Common Pre-Trends p-Values for Intensive Margin Analysis



Note: Plot of p-values of test of common pre-trends after CEM matching. The test is the joint test that all pre-period treatment effect coefficients in Equation 2.2 are equal to 0. The figure plots the p-values for all thresholds shown in Figure B.7 and Figure B.8 across all of the categories included in the News tab.

Figure B.6: Election Placebo Tests

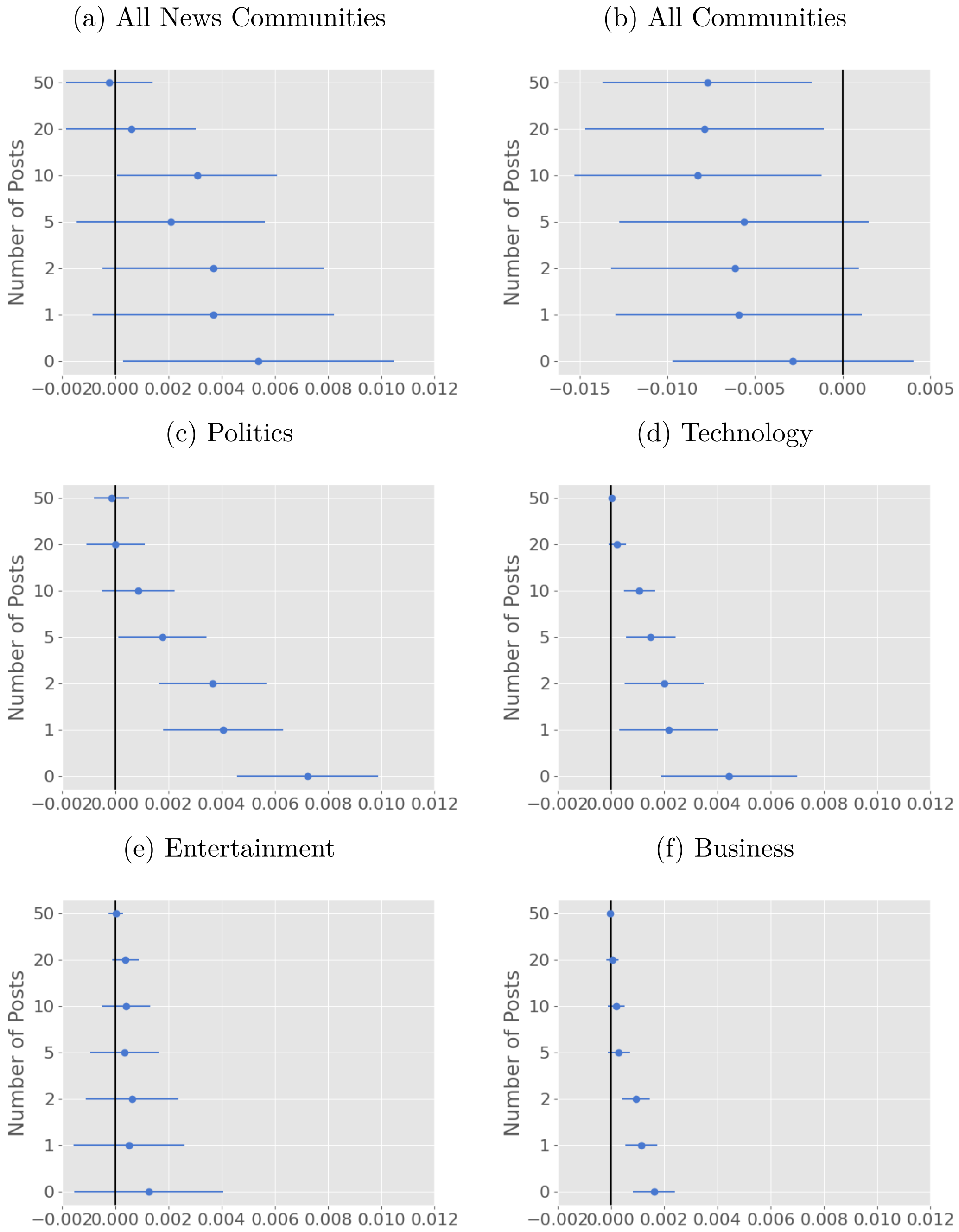


Here we show results of a placebo test on political communities not included in the News tab (red) compared to the Politics community included in the News tab (blue). We find no treatment effect on communities excluded from the News tab suggesting the effect is not confounded by the 2018 midterm election.

B.5 Robustness Checks

B.6 Effect on Intensive Margin of Engagement

Figure B.7: Treatment Effect Estimates on Probability of Posting More than k Posts


(a) All News Communities

(b) All Communities

(c) Politics

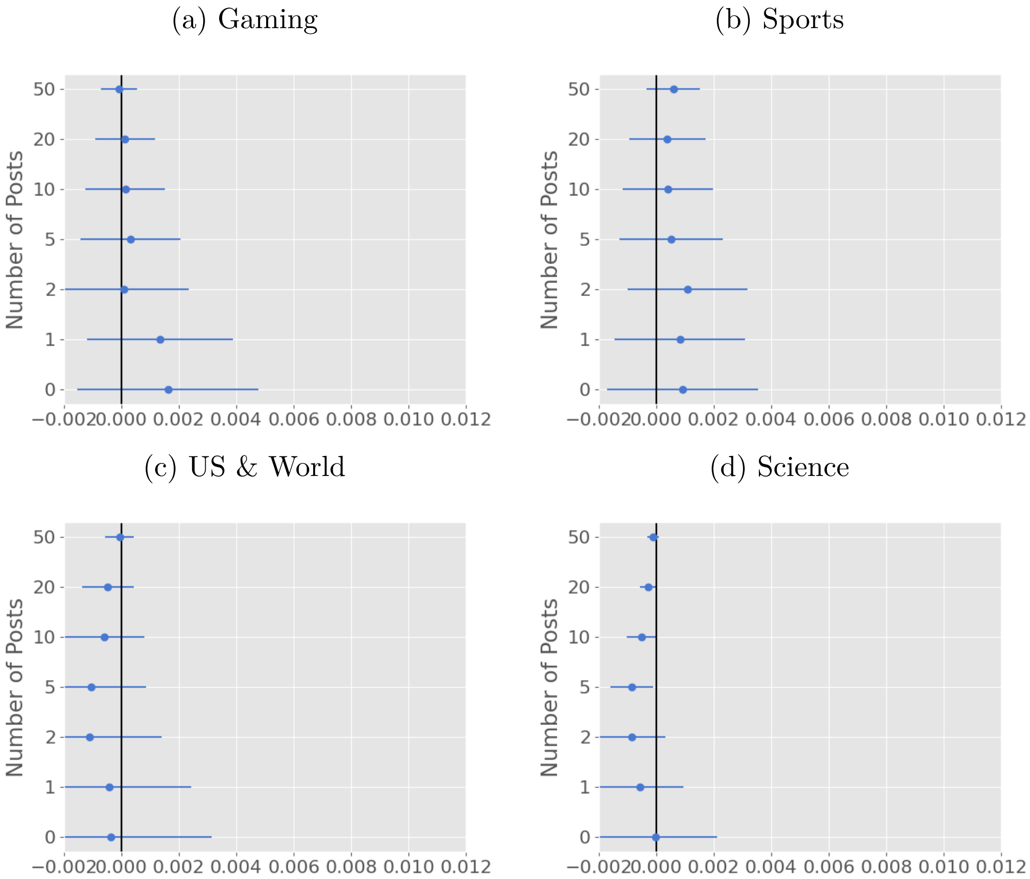
(d) Technology

(e) Entertainment

(f) Business

Note: Each plot shows the treatment effect estimate from Equation 2.1 where the outcome is an indicator if the user posted more than a threshold posts in a month (within each community). The thresholds are shown on the y-axis. Horizontal bars represent 95% confidence intervals.

Figure B.8: Treatment Effect Estimates on Probability of Posting More than k Posts



(a) Gaming

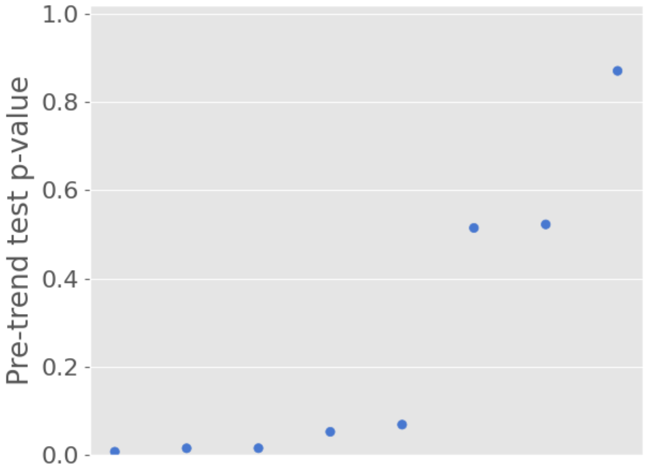
(b) Sports

(c) US & World

(d) Science

Note: Each plot shows the treatment effect estimate from Equation 2.1 where the outcome is an indicator if the user posted more than a threshold posts in a month (within each community). The thresholds are shown on the y-axis. Horizontal bars represent 95% confidence intervals.

Figure B.9: Common Pre-Trends p-values without Matching



Note: Plot of p-values of test of common pre-trends without CEM matching. The test is the joint test that all pre-period treatment effect coefficients in Equation 2.2 are equal to 0.

B.7 Analyses Without CEM Weights

Figure B.10: Treatment Effect Estimates on Probability of Posting

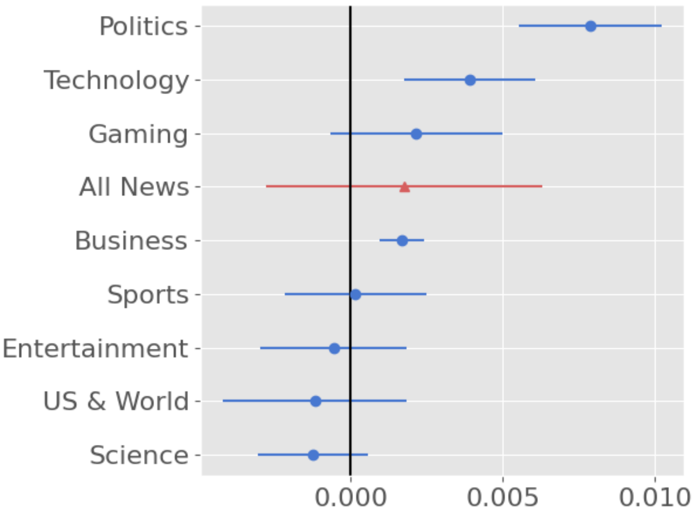
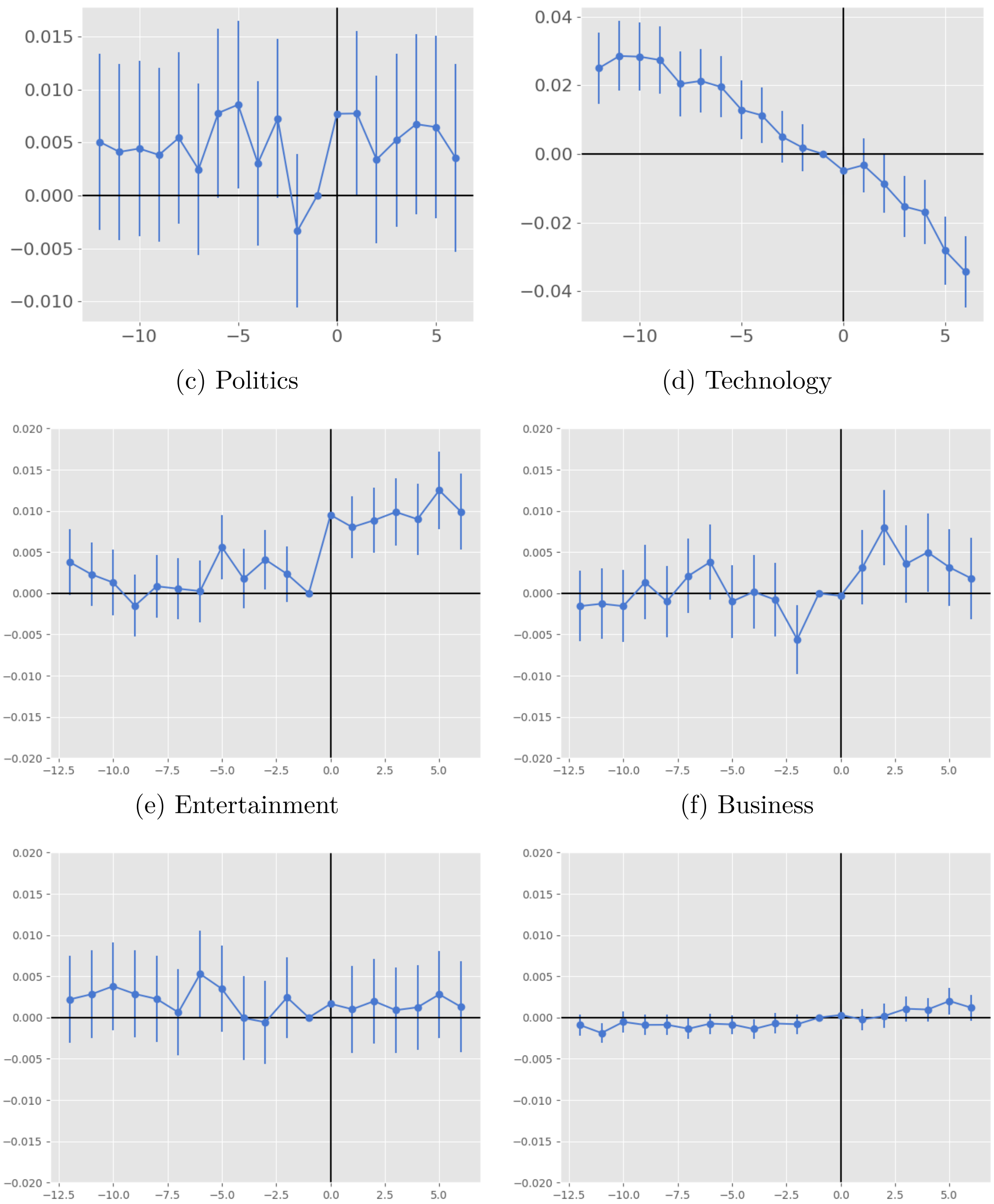


Figure B.11: Dynamic Treatment Effect Estimates on Probability of Posting without Matching



(a) All News Communities

(b) All Non-News Communities

(c) Politics

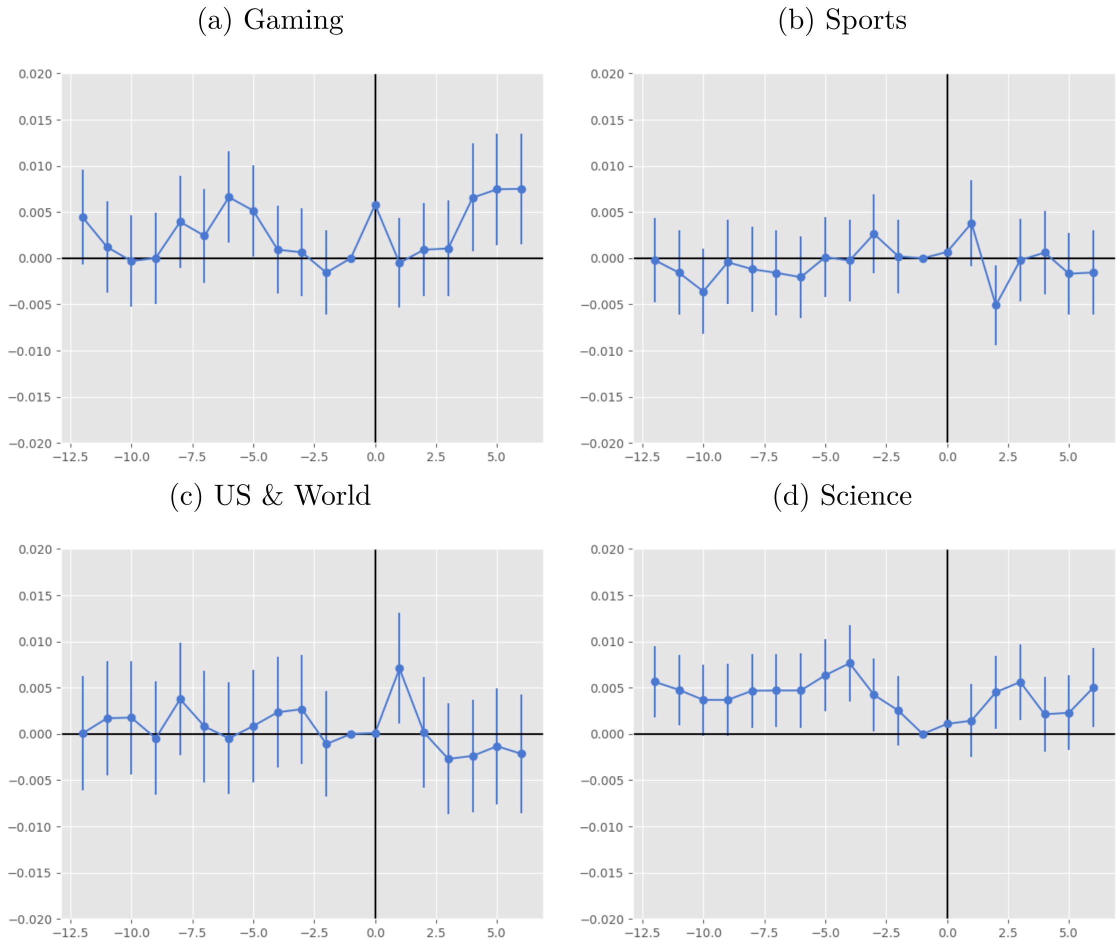
(d) Technology

(e) Entertainment

(f) Business

Note: Dynamic treatment effect estimates on probability of posting in community, estimated using Equation 2.1 without the Coarsened Exact Matching weights.

Figure B.12: Dynamic Treatment Effect Estimates on Probability of Posting without Matching


(a) Gaming

(b) Sports

(c) US & World

(d) Science

Note: Dynamic treatment effect estimates on probability of posting in community, estimated using Equation 2.1 without the Coarsened Exact Matching weights.

Appendix C

Appendix: Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology

C.1 Appendix of Proofs

C.1.1 Proof of Proposition 3.5.1

Case $b < 1$ and $d = 1$: Suppose $a^* \left(s^H; p\right) = 0$ and $a^* \left(s^A, s^H; p\right) = 1$. Equivalently, $\log c_{rel} > \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ and $\log c_{rel} \leq b \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)} + \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$. Since $b \in (0, 1)$, it must be that $\log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)} > 0$ and $\log c_{rel} < \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)} + \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ so that $a^* \left(s^A, s^H; \pi\right)$ $= 1$. Hence, if $0 = a^* \left(s^H; p\right) \neq a^* \left(s^A, s^H; p\right)$ then $a^* \left(s^A, s^H; p\right) = a^* \left(s^A, s^H; \pi\right)$ and $V \left(s^H; p\right) \leq V \left(s^A, s^H; \pi\right) = V \left(s^A, s^H; p\right)$, with strict inequality if the measure on $\left(s^A, s^H\right)$ under $\pi(\cdot)$ such that $0 = a^* \left(s^H; p\right) \neq a^* \left(s^A, s^H; \pi\right)$ is strictly positive. The proof of the case when $a^* \left(s^H; p\right) = 1$ and $a^* \left(s^A, s^H; p\right) = 0$ is analogous. If $a^* \left(s^H; p\right) = a^* \left(s^A, s^H; p\right)$ then $V \left(s^H; p\right) = V \left(s^A, s^H; p\right)$.

Case $b > 1$ and $d = 1$: If $\log c_{rel} - \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)} > 0$, then for $\log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$ $\in \left(\frac{1}{b} \log c_{rel} - \frac{1}{b} \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}, \log c_{rel} - \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}\right)$we have both $\log c_{rel} > \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$ $+ \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ and $\log c_{rel} < b \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)} + \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$. Thus, $0 = a^* \left(s^H; p\right)$ $\neq a^* \left(s^H, s^A; p\right) \neq a^* \left(s^H, s^A; \pi\right)$. An analogous argument for the case when $\log c_{rel} \leq \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ completes this case.

Case $d \neq 1$: We analyze this in two subcases.

- $(1-d)\log c_{rel} > 0$: We show that there exist values of $\left(\log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}, \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}\right)$ such that $a^*\left(s^A, s^H; p\right) = 0$, $a^*\left(s^H; p\right) = 1$, and $a^*\left(s^A, s^H; \pi\right) = 1$. Equivalently, we need to find values such that $\log c_{rel} > b\log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)} + d\log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$, $\log c_{rel} < \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ and $\log c_{rel} < \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)} + \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ if $d \neq 1$. Re-write this system as $y = \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)} - \log c_{rel}$ and $x = \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$, we need to find a solution to the system $y > 0$, $x + y > 0$ and $bx + dy < (1-d)\log c_{rel}$. Since $(1-d)\log c_{rel} > 0$, there exist small enough values of $x, y > 0$ such that the solution exists.

- $(1-d)\log c_{rel} < 0$: An argument analogous of case 1 shows that there exist values of $\left(\log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}, \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}\right)$ such that $a^*\left(s^A, s^H; p\right) = 1$, $a^*\left(s^H; p\right) = 0$, and $a^*\left(s^A, s^H; \pi\right) = 0$.

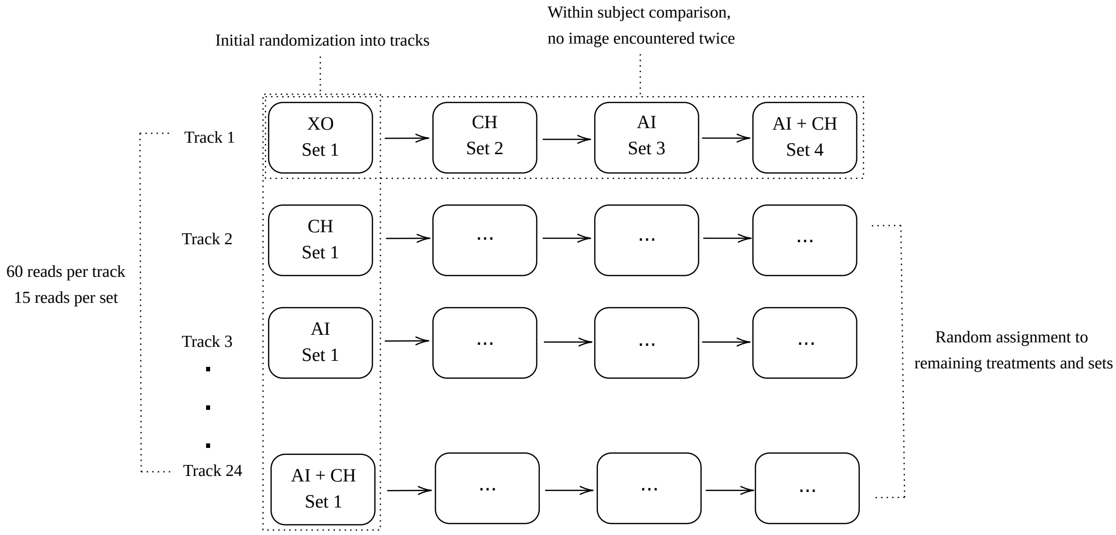
C.1.2 Proof of Proposition 3.5.2

Consider $\log c_{rel} > \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ and $\log c_{rel} < b\log \frac{\pi\left(s^A\middle|\omega=1\right)}{\pi(s^A|\omega=0)} + d\log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ so that $0 = a^*\left(s^H; p\right) \neq a^*\left(s^A, s^H; p\right) = 1$. For small enough $\log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)}$, $\log c_{rel} > \log \frac{\pi\left(s^A\middle|\omega=1,s^H\right)}{\pi(s^A|\omega=0,s^H)} + \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ so that $a^*\left(s^A, s^H; \pi\right) = 0$. The case in which $\log c_{rel} < \log \frac{\pi\left(\omega=1|s^H\right)}{\pi(\omega=0|s^H)}$ is analagous.

C.2 Appendix of Experimental Interface and Instructions

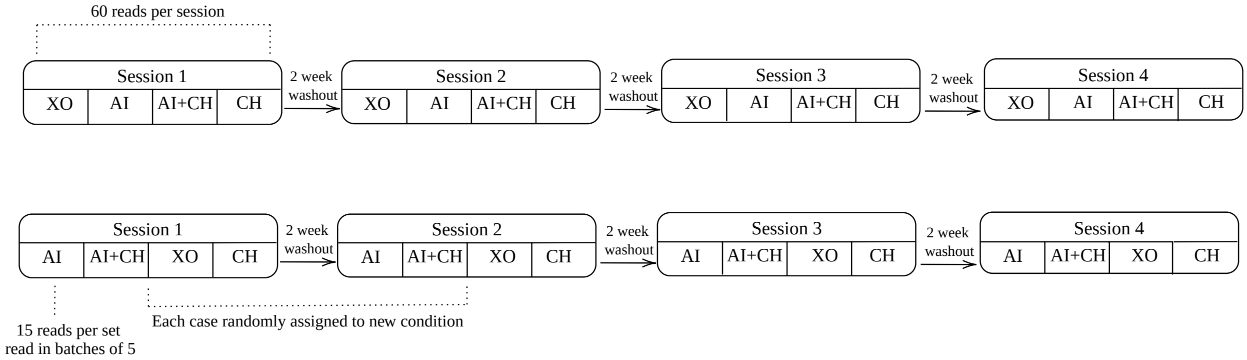
C.2.1 Design

Figure C.1: Design 1



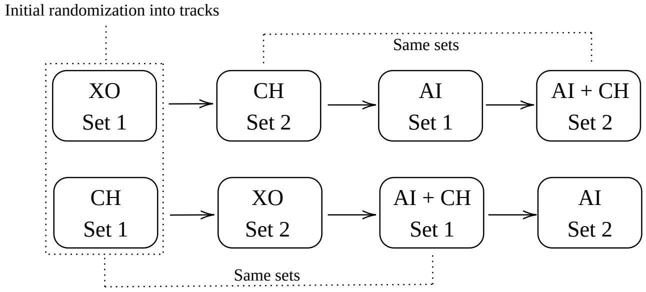
Note: In this design, radiologists are assigned to a randomized sequence of the four information environments., resulting in 24 possible tracks. Under each information environment they read 15 cases. Radiologists encounter each patient case at most once. At the beginning of the experiment every radiologist reads eight practice cases. Furthermore, a random half of the participating radiologists receive incentives for accuracy.

Figure C.2: Design 2



Note: In this design, radiologists diagnose 60 patient cases each under the four information environments. Radiologists read every case under every information environment across four sessions, separated by a washout period. Each case is only encountered once per session and to ensure that radiologists do not recall their/AI predictions from previous reads of the same cases, we ensure a minimum two-week washout period between subsequent sessions. Within every experimental session radiologists therefore read 15 under each information environment. The randomization occurs at the track-level where every track has a different sequence of the information environments. (Example tracks shown here.)

Figure C.3: Design 3



Note: In this design, radiologists diagnose 50 cases, first without and then with AI assistance. Clinical history is randomly provided in either the first or second half of images forming the basis of the randomization. The cases diagnosed with and without clinical history are different.

C.2.2 Instructions

Below are the instructions the subjects received along with the interface-based treatment. Comments on the instructions are provided in italics and were not seen by subjects.

Instructions

You are about to participate in a study on medical decision making. You may pause the study at any time. To resume, revisit the link you were given and your progress will have been saved.

We will present you with adult patients with potential thoracic pathologies. These patients will be presented under the following four scenarios:

1. Only a chest X-ray is shown.
2. An X-ray is accompanied with additional information about the clinical history.
3. An X-ray is shown along with Artificial Intelligence (AI) support. This AI tool is described in further detail below.
4. An X-ray is shown along with both additional information on clinical history and the AI support.

The patients are randomly assigned to each of these scenarios. That is, availability of clinical history and/or AI support is unrelated to the patient.

Clinical History: includes available lab results or indications by the treating physician, if any.

AI support: This tool uses only the X-ray image to predict the probability of each potential pathology of interest. The tool is based on state-of-the-art machine learning algorithms developed by a leading team of researchers at Stanford University.

Responses

For each patient and pathology, we will ask for both an assessment and a treatment decision:

1. We will first ask for your assessment of the probability that each condition is present in a patient. **Please consider all pathologies and findings that would be relevant in a radiology report for the patient. You should express your uncertainty about the presence of one or many conditions by appropriately choosing the probability.** Note that it is possible that the patient has multiple such conditions or none of them.

2. If you determine that a pathology may be present, we may ask you to rate the severity and/or extent of the disease on a scale.

3. Finally, when relevant we will ask whether you would recommend treatment or follow-up according to the clinical standard of care if you determine that the pathology may be present. The first two responses are diagnostic while the third is a clinical decision. We are aware that a single physician or radiologist typically does not perform both tasks. However, for this study, we ask that you respond to the best of your ability in both of these roles.

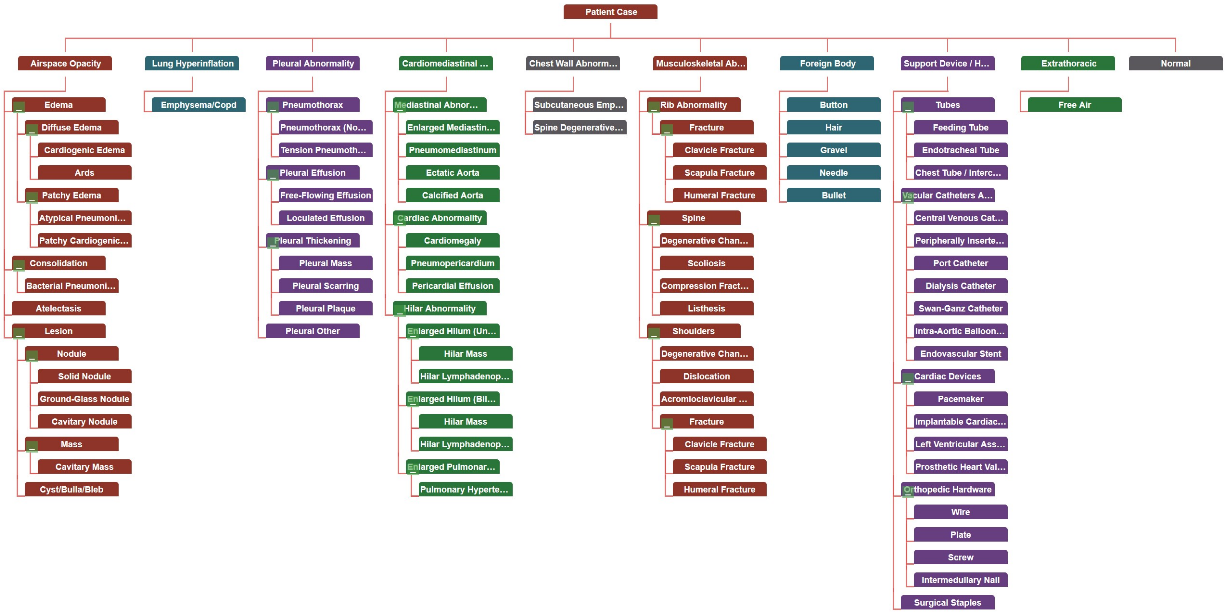
Browser Compatibility

This platform supports desktop versions of Chrome, Firefox, and Edge. Important features on non-supported browsers (including Safari) are missing and we discourage their use for this experiment. In addition, the platform does not support any mobile devices and the platform will perform poorly on mobile. If you encounter any issues during the experiment, please send an email to DiagnosticAI@mit.edu and we will follow-up quickly.

Hierarchy

The interface uses a hierarchy to categorize various thoracic conditions. It will be useful to familiarize yourself with this hierarchy before you start, but you may also revisit the hierarchy at any time throughout the experiment by clicking the help tab in the upper right corner. *[The probability for the sub-pathologies is required only if the parent pathology prevalence is greater than 10%.]*

Figure C.4: Pathology hierarchy



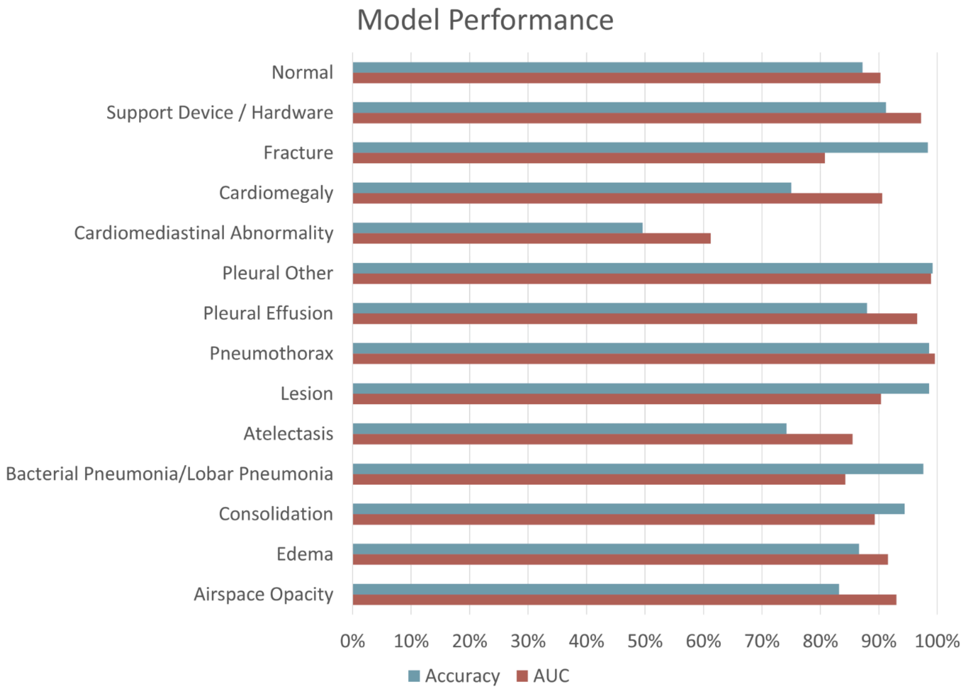
AI Support Tool

The AI support tool that is provided uses only the X-ray image to predict the probability of each potential pathology of interest. The tool is based on state-of-the-art machine learning algorithms developed by a leading team of researchers at Stanford University. The tool is trained only on X-ray images, meaning it does not incorporate the clinical history of the patients.

Performance of the AI Support

The AI tool is described in Irvin et al. [2019], which showed the AI tool performed at or near expert levels across the pathologies studied. Below we plot two measures of performance of the AI tool. We plot in blue the accuracy of the tool, defined as the share of cases correctly diagnosed when treating false positives and false negatives equally. In red, we plot the Area Under the ROC curve (AUC), which is another measure of AI classification performance. The AUC is a number between 0 and 100%, with numbers close to 100% representing better algorithm performance. The AUC is equal to the probability that a randomly chosen positive case is ranked higher than a randomly chosen negative case.

Figure C.5: Performance of AI tool

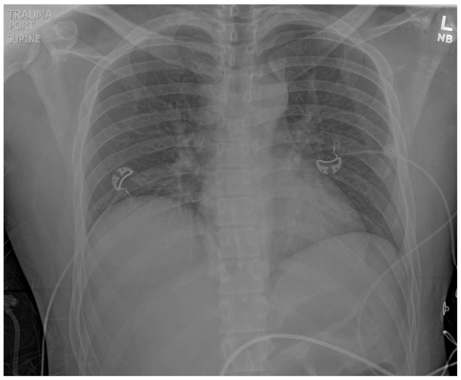


Example Images

Below are 50 example images with the associated AI tool predictions. These images are randomly chosen to allow you to familiarize yourself with the AI support tool and its accuracy. *[Here we only provide two out of the 50 images. Notice that these assessments need not sum to 100% as a case can have more than one pathology. The sum of assessments among pathologies that are nested within a top-level pathology also may be less than the top-level pathology's assessment as a case could have the top-level pathology but none of the child pathologies with an AI prediction.]*

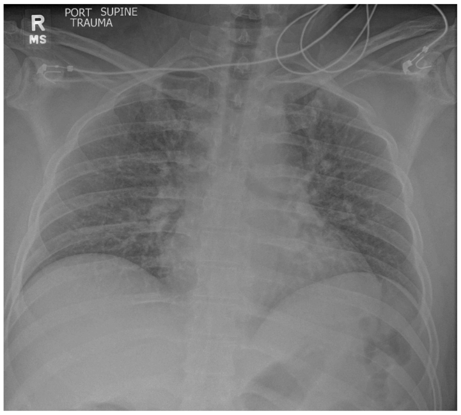
Figure C.6: Example images

Example 1



Pathology	AI Prediction
Airspace Opacity	16%
• Edema	7%
• Consolidation	3%
◦ Bacterial Pneumonia/Lobar Pneumonia	3%
• Atelectasis	9%
• Lesion	4%
Pleural Abnormality	
• Pneumothorax	7%
• Pleural Effusion	1%
• Pleural Other	0%
Cardiomediastinal Abnormality	14%
◦ Cardiomegaly	1%
Musculoskeletal Abnormality	
◦ Fracture	9%
Support Device / Hardware	12%
Normal	47%

Example 2



Pathology	AI Prediction
Airspace Opacity	42%
• Edema	42%
• Consolidation	14%
◦ Bacterial Pneumonia/Lobar Pneumonia	14%
• Atelectasis	5%
• Lesion	5%
Pleural Abnormality	
• Pneumothorax	3%
• Pleural Effusion	0%
• Pleural Other	1%
Cardiomediastinal Abnormality	16%
◦ Cardiomegaly	5%
Musculoskeletal Abnormality	
◦ Fracture	9%
Support Device / Hardware	3%
Normal	21%

Demonstration

The brief video below walks you through the interface and a few examples. *[At this stage participants saw an instructional video which can be found here.]*

Consent

You have been asked to participate in a study conducted by researchers from the Massachusetts Institute of Technology (M.I.T.) and Harvard University.

The information below provides a summary of the research. Your participation in this research is voluntary and you can withdraw at any time.

1. Study procedure: We will ask you to examine a number of chest x-rays. We will vary both the amount of information provided about the patient and the availability of an AI support tool.

2. Potential Risks & Benefits: There are no foreseeable risks associated with this study and you will receive no direct benefit from participating.

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise.

Privacy & Confidentiality

The only people who will know that you are a research subject are members of the research team which might include outside collaborators not affiliated with MIT. No identifiable information about you, or provided by you during the research, will be disclosed to others without your written permission, except: if necessary to protect your rights or welfare, or if required by law. In addition, your information may be reviewed by authorized MIT representatives to ensure compliance with MIT policies and procedures.

When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity.

Questions

If you have any questions or concerns about the research, please feel free to contact us directly at diagnosticAI@mit.edu.

Your Rights

You are not waiving any legal claims, rights, or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

I understand the procedures described above. By clicking next, I am acknowledging my questions have been answered to my satisfaction, and I agree to participate in this study.

Interface questions

[Each of these questions has a true or false response which was entered through a radio button. Participants are not able to start the experiment without answering each question correctly.]

Before beginning the experiment, we would like to confirm a few facts through the following comprehension questions. Please answer True or False to the following questions.

1) The algorithm's prediction is based on information from both the X-ray scan as well as the clinical history.

2) When the algorithm does not show a prediction, it is because the algorithm thinks the pathology is not present.

3) The follow-up decision refers to any treatment or additional diagnostic procedures that one would conduct based on the findings of the report.

4) Two patients with the same probability score for a condition ought to always receive the same "follow-up" recommendation.

5) When a condition at a higher level of the hierarchy receives a less than ten percent chance of being present then all the lower level conditions within this branch automatically receive a zero probability of being present.

6) If the algorithm says that the probability of a pathology is present with 80% probability, it means that the AI predicts 80 cases out of 100 have the pathology present.

7) Suppose your assessment is that the patient definitely has either edema or consolidation, and you believe that edema is twice as likely as consolidation. Then you would assign 66.67% to edema and 33.33% to consolidation.

8) I should only indicate pathologies and findings that would be relevant in a radiology report for the patient.

Interface

Figure C.7 is an example of the clinical history indications available to the participating radiologists under the relevant treatment condition. The thoroughness of the information varies across available information for every patient. Some examples of varying clinical history information are:

1. 68 years of age, Female, chest pain

2. Unknown age, Unknown, trauma

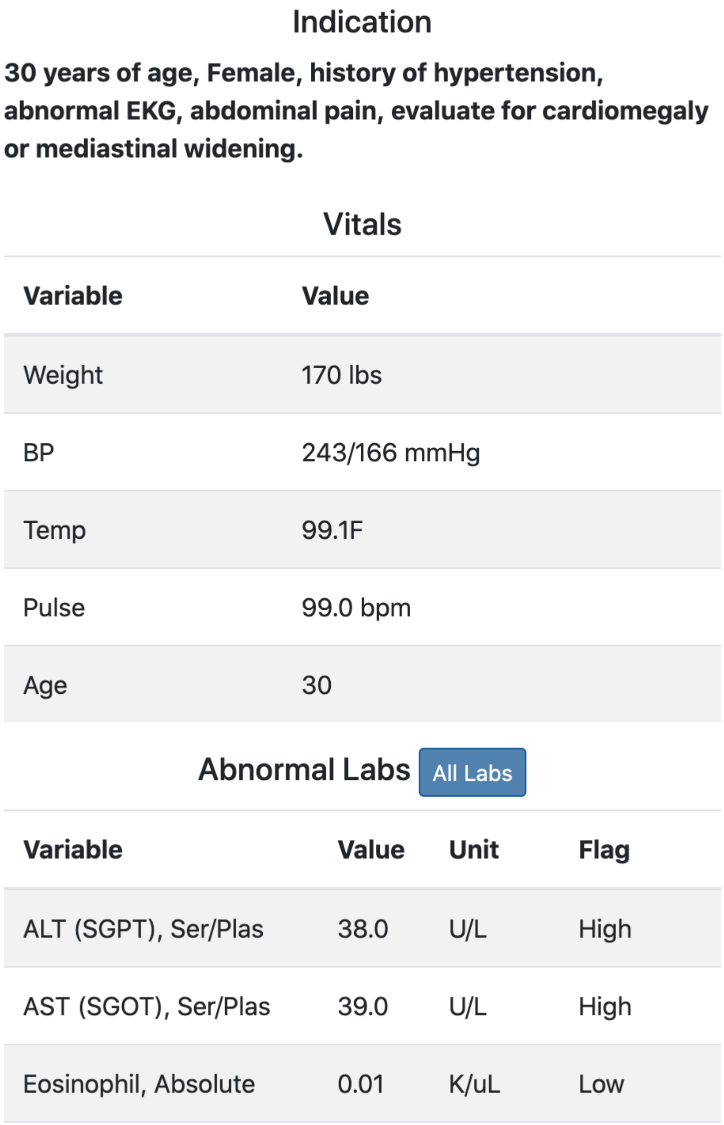
3. 55 years of age, Male, Order History: Relevant PMH gastroparesis. Presents with vomiting, retching chest discomfort for a duration of today. Concern for PTX, perforated viscus, pneumomediastinum

4. 74 years of age, Female, s/p unwitnessed fall, r/o rib fx, pna or effusion

5. Trauma

6. 56 years of age, Male, S/P ICD/ Pacemaker insertion / Complete X-ray without lifting arms above shoulders..

Figure C.7: Clinical history information



Indication

30 years of age, Female, history of hypertension, abnormal EKG, abdominal pain, evaluate for cardiomegaly or mediastinal widening.

Vitals

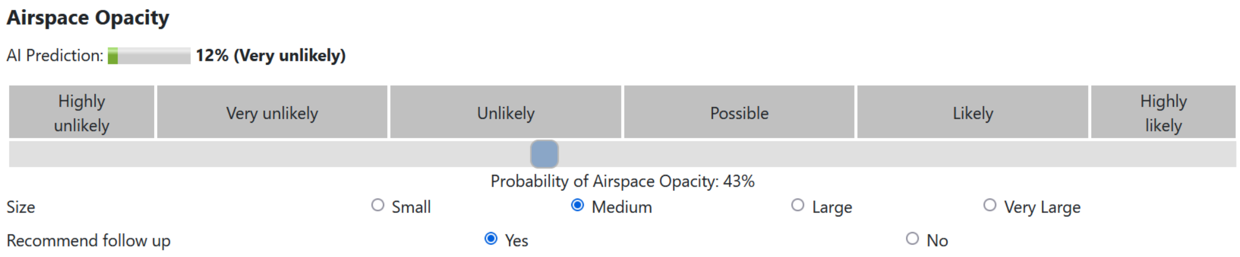
Variable	Value
Weight	170 lbs
BP	243/166 mmHg
Temp	99.1F
Pulse	99.0 bpm
Age	30

Abnormal Labs

Variable	Value	Unit	Flag
ALT (SGPT), Ser/Plas	38.0	U/L	High
AST (SGOT), Ser/Plas	39.0	U/L	High
Eosinophil, Absolute	0.01	K/uL	Low

Note: The clinical history information environment in the experiment had information on patient indications, vitals, and abnormal labs.

Figure C.8: Interface slider



Airspace Opacity

AI Prediction: **12% (Very unlikely)**

Probability of Airspace Opacity: 43%

Size ○ Small ● Medium ○ Large ○ Very Large

Recommend follow up ● Yes ○ No

Note: The participants use the slider to indicate the probability of a pathology being present for a given patient based on the treatment offered. For prevalence greater than 10% the participants are required to indicate the prevalence of a sub-pathology (if it exists) and whether a follow-up is recommended.

C.2.3 Additional Details on the AI Algorithm

The training data is a set of tuples of images and labels. These training datasets typically rely on human input to assign the labels, which indicate whether or not a specific pattern or object is present in the image. Training is conducted through stochastic gradient descent. These algorithms build on the nested structure of the neural net to compute gradients computationally efficiently via the chain rule. Each training step is performed on a small batch of data so that the algorithm does not have to consider the entire dataset for each optimization step. After each round of optimization on the training set, the model performance is assessed through predictions on a hold-out validation sample. Most humans are able to recognize cars, pedestrians, and traffic lights, which means that training datasets for common classification tasks are easy to come by. The same is not true for medical imaging. Classifying disease based on X-rays, CT scans, and retina scans requires the input of highly trained experts. Recently, several researchers have released large training datasets of medical images with disease labels that are extracted from written clinical descriptions [Irvin et al., 2019]. The neural net has a DenseNet121 architecture. A DenseNet is a type of convolutional neural network that utilizes dense connections between layers through Dense Blocks; in these blocks, we connect all layers with matching feature-map sizes directly with each other. Images are supplied in a standardized format of 320 × 320 pixels. For optimization the researchers use the Adam optimizer with default b-parameters of $b1 = 0.9$, $b2 = 0.999$ and learning rate $1 \times 10{-}4$. The batch size is fixed at 16 images. The training is performed for 3 epochs. The full training procedure is described in [Irvin et al., 2019].

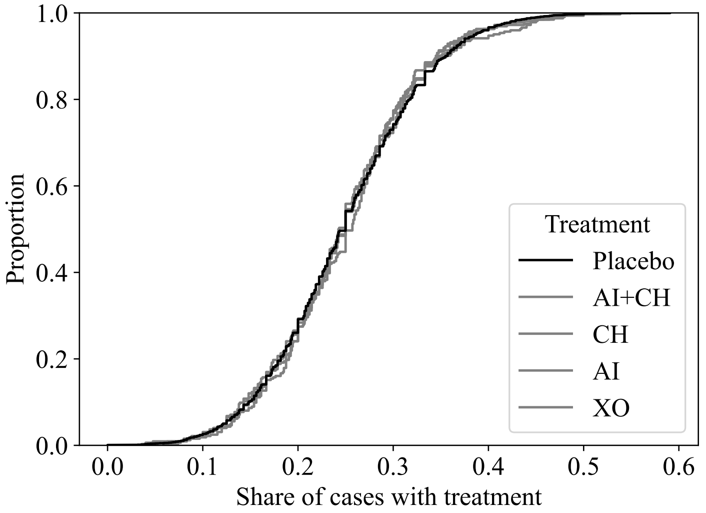
C.3 Data Appendix

C.3.1 Balance Tests

We verify that the randomization occurred as expected through various balance and randomization tests. Figure C.9 plots the distribution of treatment probabilities by patient-case in Design 1. We also plot a placebo distribution that samples from the null distribution to support the claim that the randomization occurred as expected. To test this formally, we present balance tests for Design 1 and Design 2 in Table C.1 and Table C.2 , respectively.[1] For these balance tests, we calculate the average covariates across the four treatment arms and report p-values from the test of the joint null that the four means are equal. For Design 2, these are done within sessions as patients are balanced by design across all sessions.

[1] Design 3 is balanced by design, as each radiologist reads the same cases with and without AI assistance.

Figure C.9: Distribution of patient treatment probabilities in design 1



Note: The cumulative distribution functions of patient treatment probabilities by treatment for design 1. The placebo distribution is calculated based on 100,000 draws from the null distribution. For each draw from the null distribution, we sample the number of reads the case receives from the empirical distribution and then draw the number of treatments from a binomial distribution with probability 1/4.

Table C.1: Covariate balance in design 1

	Control	CH	AI	AI x CH	p-value
s_A	0.309	0.301	0.310	0.306	0.310
Airspace Opacity	0.163	0.149	0.166	0.159	0.404
Cardiomediastinal Abnormality	0.131	0.130	0.138	0.131	0.832
Support Device Hardware	0.176	0.169	0.176	0.190	0.292
Abnormal	0.187	0.179	0.195	0.189	0.545
Weight	185.24	185.87	185.20	185.17	0.942
Temp	99.02	99.04	99.05	99.06	0.230
Pulse	92.26	92.72	92.55	92.92	0.074
Age	56.80	56.55	56.42	56.87	0.858
Number Labs	34.61	34.23	34.54	34.29	0.372
Number Flagged Labs	5.907	5.862	6.061	6.053	0.349
Female	0.416	0.409	0.389	0.388	0.101

Note: Balance tests of patient covariates for patients assigned to the four treatments in Design 1. Missing clinical history variables are mean-imputed. The p-values come from the joint test the mean covariates are equal across the four treatments.

Table C.2: Covariate balance in design 2

	Session 1	Session 2	Session 3	Session 4
s_A	0.381	0.625	0.381	0.447
Airspace Opacity	0.243	0.368	0.141	0.483
Cardiomediastinal Abnormality	0.164	0.834	0.088	0.716
Support Device Hardware	0.760	0.770	0.714	0.794
Abnormal	0.265	0.624	0.722	0.330
Weight	0.461	0.597	0.878	0.735
Temp	0.107	0.245	0.437	0.654
Pulse	0.242	0.578	0.764	0.772
Age	0.559	0.220	0.082	0.898
Number Labs	0.075	0.348	0.581	0.768
Number Flagged Labs	0.297	0.189	0.935	0.738
Female	0.067	0.052	0.225	0.075

Note: Balance test p-values that the covariate means are equal across the four treatments within each session (column). Missing clinical history variables are mean-imputed.

C.3.2 Quality of Diagnostic Standard

Here, we summarize evidence that the diagnostic standard measure we construct is high quality and robust to various decisions an analyst could make. Recall that the preferred diagnostic standard used throughout the paper is defined using the reads of five board-certified radiologists from Mount Sinai, who each read all 324 patient cases in the study in a random order. For each pathology, we aggregate these reports into the diagnostic standard for a patient case i as

$$\omega_i = 1\left[\sum_{r=1}^{5} \frac{\pi_r(\omega_i = 1|s^E_{i,r})}{5} > \frac{1}{2}\right]$$

where we suppress the pathology index for simplicity and r indexes the radiologist. This method of aggregating reports is robust to certain types of measurement error and dependence across reports as discussed in Wallsten and Diederich [2001]. Table C.3 contains summary statistics for the diagnostic standard created using the Mount Sinai radiologists and a leave-one-out internal diagnostic standard calculated using the reads collected during the experiment under the treatment arm with clinical history but no AI assistance. Table C.4 contains additional summary statistics for the five Mount Sinai diagnostic standard labelers, including their average time and number of clicks. We also show the average agreement of the labels with the original radiologist's read. Taken together, these analyses demonstrate, for the majority of cases, that the diagnostic standard labelers agree with the assessment of the radiologist who originally read the report in a clinical setting and we can reject that the average probability assessment is equal to 0.5 at the 5% level. Moreover, in Section C.3.5 we show that our results are robust to many different methods of calculating diagnostic standard, including using the experiment leave-one-out diagnostic standard and various aggregation methods of the Mount Sinai reports.

Table C.3: Diagnostic standard quality

	Prevalence		Share Rejecting 0.5		Average Number of Rads	
	Sinai	Experiment	Sinai	Experiment	Sinai	Experiment
Top-Level with AI	0.147	0.110	0.696	0.795	5.00	16.22
Pooled with AI	0.043	0.028	0.892	0.940	5.00	16.22
Abnormal	0.194	0.506	0.583	0.565	5.00	16.22
All Pathologies	0.013	0.009	0.953	0.980	5.00	16.22

Note: For each of the pre-registered pathology groups, this table shows the average prevalence, the share of cases where we can reject that $\sum_{r=1}^{R} \frac{\pi_r(\omega_i=1|s^E_{i,r})}{R} = 0.5$ at the 5% level, and the average number of reads per case for both the Mount Sinai diagnostic standard and the experiment leave-one-out diagnostic standard.

Table C.4: Diagnostic standard effort

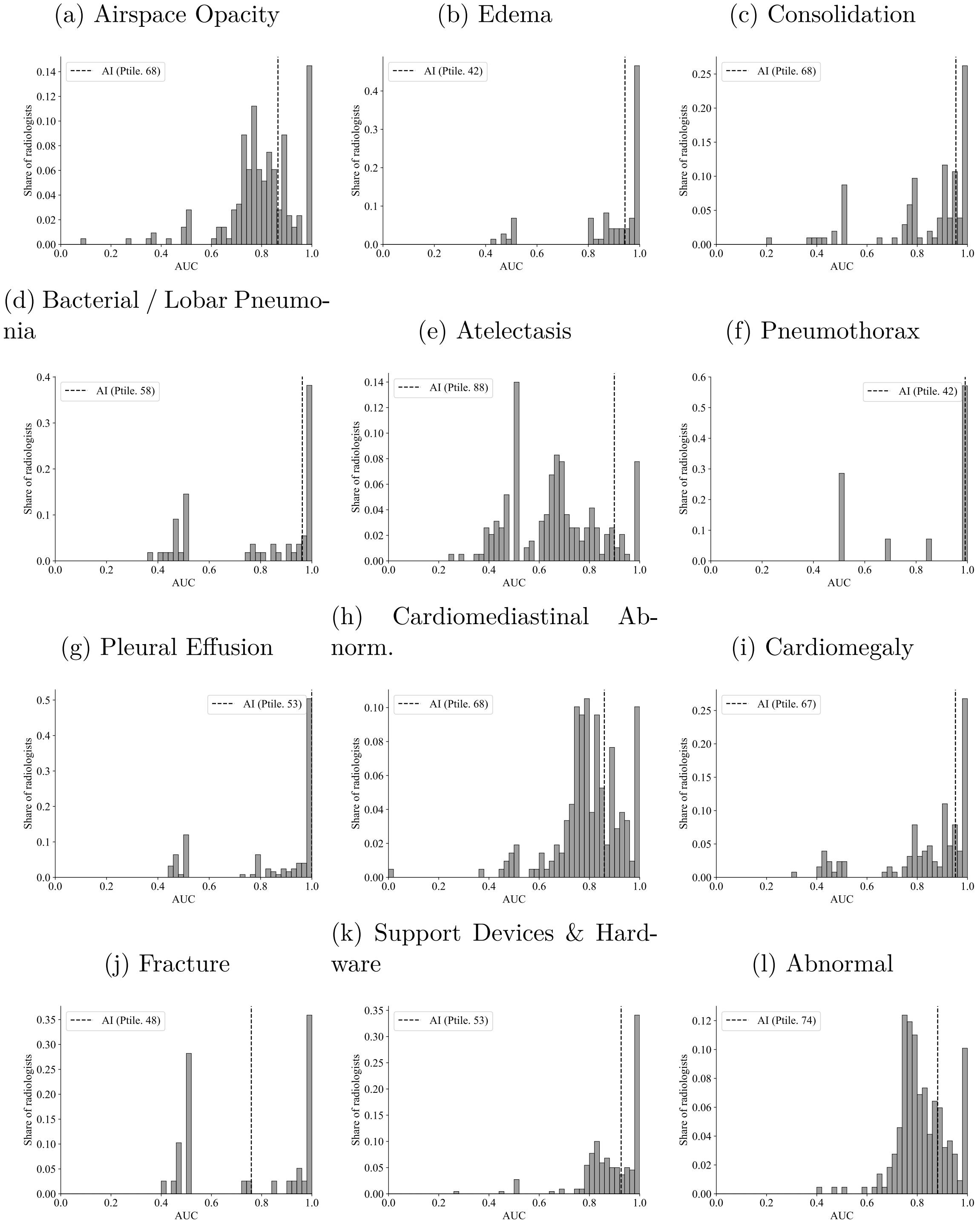
	Active Time		Clicks		Agreement with Original
	Mean	SD	Mean	SD	
0	77.24	42.78	34.30	17.93	0.868
1	76.44	54.71	32.30	18.24	0.851
2	25.55	30.34	10.84	12.29	0.876
3	79.94	80.22	21.79	20.59	0.866
4	112.96	82.96	26.14	20.12	0.863

Note: For each of the five Mount Sinai radiologists we compute the average and standard deviation of time spent per case and the number of clicks per case. In addition, we compute the average agreement with the original read as labeled by the CheXbert algorithm.

C.3.3 Performance Distributions by Pathology

This section presents distributions for two different accuracy measures for radiologists and the AI across different pathology groups. These figures allow for a comparison between the accuracy of the AI relative to the mean radiologist.

Figure C.10: AUROC


(a) Airspace Opacity

(b) Edema

(c) Consolidation

(d) Bacterial / Lobar Pneumonia

(e) Atelectasis

(f) Pneumothorax

(g) Pleural Effusion

(h) Cardiomediastinal Abnorm.

(i) Cardiomegaly

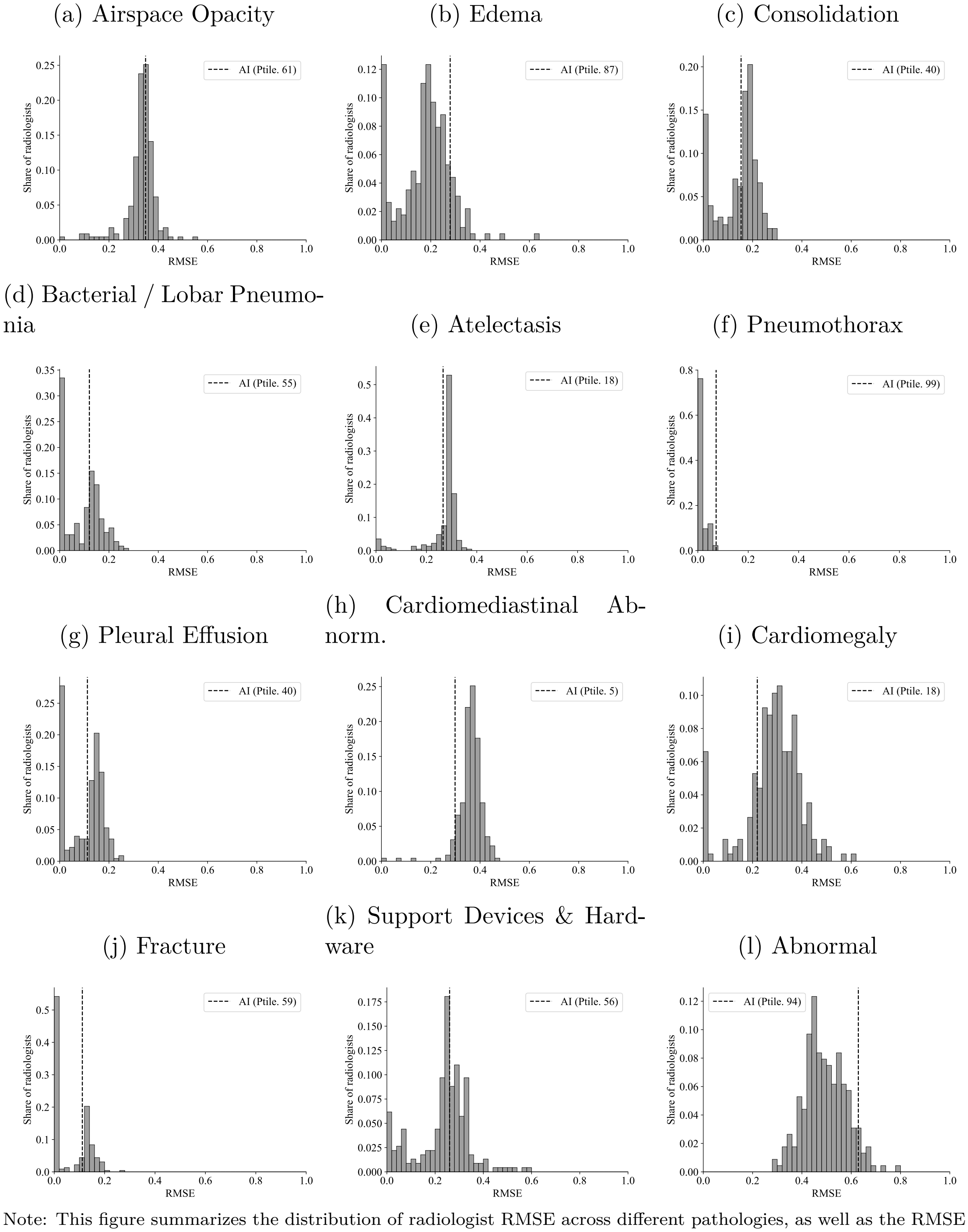
(j) Fracture

(k) Support Devices & Hardware

(l) Abnormal

Note: This figure summarizes the distribution of radiologist AUROCs across different pathologies, as well as the AUROC of the AI algorithm for the corresponding pathology. Only the cases where contextual history information is available for the radiologist but not the AI prediction were considered. AUROC is only defined for radiologists who encounter some positive cases.

Figure C.11: RMSE


(a) Airspace Opacity

(b) Edema

(c) Consolidation

(d) Bacterial / Lobar Pneumonia

(e) Atelectasis

(f) Pneumothorax

(g) Pleural Effusion

(h) Cardiomediastinal Abnorm.

(i) Cardiomegaly

(j) Fracture

(k) Support Devices & Hardware

(l) Abnormal

Note: This figure summarizes the distribution of radiologist RMSE across different pathologies, as well as the RMSE of the AI algorithm for the corresponding pathology. Only the cases where contextual history information is available for the radiologist but not the AI prediction were considered.

C.3.4 Comparison of Radiologists to Original Reads

The reports from the radiologists who originally read the patient cases included in our sample were classified as positive/negative/uncertain for each pathology using AI predictions generated by the CheXbert algorithm described in Smit et al. [2020]. We compare the accuracy of the original reads relative to the diagnostic standard with the radiologists in our sample under the treatment arm with clinical history and no AI assistance. We do this for each pathology by converting the probability reports elicited during the experiment to positive/negative assessments, where positive is defined as having a probability greater than 50%. We convert the CheXbert labels to positive/negative assessments by including the uncertain cases as positive.[2] We then calculate the accuracy of the experiment reads and the CheXbert labels for groups of pathologies focused on in this study and test the null hypothesis that the accuracy of the radiologists is the same. The results of this analysis are in Table C.5, and those for when treating uncertain cases negative are in Table C.6.

Table C.5: Comparing experiment assessments to original reads

	Top-Level with AI	Pooled with AI	Abnormal
	(1)	(2)	(3)
Experiment	-0.000	-0.004	-0.086
	(0.016)	(0.006)	(0.025)
Constant	0.194	0.090	0.466
	(0.016)	(0.006)	(0.028)
Observations	11128	61204	5564
R-Squared	0.000	0.000	0.002

$^{*}p < 0.1$; $^{**}p < 0.05$, $^{***}p < 0.01$

Note: Regression of indicator equal to one if binarized assessment is equal to the diagnostic standard from both the original reads and experiment reads onto a constant and an indicator equal to one if the radiologist was in the experiment. Standard errors are clustered at the patient-case level.

[2] For all pathologies but bacterial pneumonia and atelectasis, fewer than 5% of patients have uncertain cases. For abnormal and all of the top-level pathologies with AI, there are no cases with uncertain labels.

Table C.6: Comparing experiment to original reads: uncertain as not present

	Top-Level with AI	Pooled with AI	Abnormal
	(1)	(2)	(3)
Experiment	-0.000	0.021	-0.086
	(0.016)	(0.004)	(0.025)
Constant	0.194	0.065	0.466
	(0.016)	(0.005)	(0.028)
Observations	11128	61204	5564
R-Squared	0.000	0.000	0.002

*$p < 0.1$; **$p < 0.05$, ***$p < 0.01$

Note: Regression of indicator equal to one if binarized assessment is equal to the diagnostic standard from both the original reads and experiment reads onto a constant and an indicator equal to one if the radiologist was in the experiment. Standard errors are clustered at the patient-case level.

C.3.5 Robustness

In this section, we show the robustness of the results from Section 3.4.2. We first present a table version of the results presented in figure 3.2 including various combinations of fixed effects (tables C.7-C.9). We next present robustness of the results in Section 3.4.2 by experiment design and by definition of the diagnostic standard. In addition, we test for order effects and test the impact of incentives.

Table C.7: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**			
	All Designs	Design 1	All Designs	Design 1	All Designs		Design 1	
					Active Time	Clicks	Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI × CH	0.002	0.002	0.001	−0.003	−1.21	0.07	−0.89	0.01
	(0.003)	(0.005)	(0.005)	(0.010)	(3.60)	(0.76)	(5.94)	(1.29)
AI	−0.040	−0.041	0.003	0.004	5.94	1.22	4.94	1.27
	(0.004)	(0.005)	(0.004)	(0.007)	(2.44)	(0.54)	(4.06)	(0.89)
CH	−0.001	−0.003	−0.009	−0.010	8.12	0.24	8.15	0.26
	(0.002)	(0.004)	(0.004)	(0.007)	(2.50)	(0.52)	(4.15)	(0.89)
Control Mean	0.212	0.222	0.226	0.223	154.32	42.65	154.47	38.88
	(0.006)	(0.007)	(0.010)	(0.011)	(4.18)	(1.15)	(6.05)	(1.36)
Pathology FE	Yes	Yes	Yes	Yes	–	–	–	–
Radiologist FE	No	No	No	No	No	No	No	No
Case FE	No	No	No	No	No	No	No	No
Observations	41920	19080	41920	19080	17455	17455	9538	9538

Note: This table summarizes the average treatment effects (ATE) of different information environments on the absolute value of the difference between the radiologist probability and AI probability (column (1) and (2)); absolute value of the difference between the radiologist probability and the diagnostic standard (columns (3) and (4)); and radiologists' effort measured in terms of active time and clicks (columns (5), (6), (7) and (8)). We either pool across all designs (All Designs) or condition on only design 1. Results on effort measure excludes five patient-cases with unaccounted time measure, and observations from design 3 because of learning effects in this set-up. Active time is winsorized to the 95th percentile. The results are for the two top-level pathologies with AI predictions, airspace opacity and cardiomediastinal abnormality. Standard errors are two-way clustered at the radiologist and patient-case level in parenthesis. Robustness by design can be found in section C.3.5.

Table C.8: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**			
	All Designs	Design 1	All Designs	Design 1	All Designs		Design 1	
					Active Time	Clicks	Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI × CH	0.002	0.002	0.001	−0.003	−1.22	0.07	−0.90	0.00
	(0.003)	(0.005)	(0.005)	(0.010)	(3.60)	(0.76)	(5.95)	(1.29)
AI	−0.040	−0.041	0.003	0.004	5.94	1.23	4.94	1.27
	(0.003)	(0.005)	(0.004)	(0.007)	(2.44)	(0.54)	(4.06)	(0.89)
CH	−0.001	−0.003	−0.009	−0.010	8.12	0.24	8.17	0.26
	(0.002)	(0.004)	(0.004)	(0.007)	(2.50)	(0.52)	(4.15)	(0.89)
Control Mean	0.212	0.222	0.226	0.223	154.31	42.64	154.46	38.88
	(0.005)	(0.006)	(0.009)	(0.010)	(1.44)	(0.31)	(2.35)	(0.51)
Pathology FE	Yes	Yes	Yes	Yes	-	-	-	-
Radiologist FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case FE	No	No	No	No	No	No	No	No
Observations	41920	19080	41920	19080	17455	17455	9538	9538

Note: This table summarizes the average treatment effects (ATE) of different information environments on the absolute value of the difference between the radiologist probability and AI probability (column (1) and (2)); absolute value of the difference between the radiologist probability and the diagnostic standard (columns (3) and (4)); and radiologists' effort measured in terms of active time and clicks (columns (5), (6), (7) and (8)). We either pool across all designs (All Designs) or condition on only design 1. Results on effort measure excludes five patient-cases with unaccounted time measure, and observations from design 3 because of learning effects in this set-up. Active time is winsorized to the 95th percentile. The results are for the two top-level pathologies with AI predictions, airspace opacity and cardiomediastinal abnormality. Standard errors are two-way clustered at the radiologist and patient-case level in parenthesis. Robustness by design can be found in section C.3.5.

Table C.9: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**			
	All Designs	Design 1	All Designs	Design 1	All Designs		Design 1	
					Active Time	Clicks	Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI × CH	0.002	0.004	0.001	−0.003	−1.18	0.01	−0.58	−0.05
	(0.003)	(0.005)	(0.004)	(0.007)	(3.48)	(0.69)	(5.74)	(1.16)
AI	−0.041	−0.043	0.002	0.002	5.55	1.11	3.70	1.00
	(0.003)	(0.004)	(0.004)	(0.006)	(2.30)	(0.49)	(3.79)	(0.79)
CH	−0.002	−0.003	−0.008	−0.005	8.38	0.41	8.65	0.51
	(0.002)	(0.003)	(0.003)	(0.005)	(2.43)	(0.46)	(4.05)	(0.78)
Control Mean	0.213	0.222	0.226	0.222	154.37	42.63	154.76	38.91
	(0.002)	(0.002)	(0.002)	(0.003)	(1.39)	(0.27)	(2.26)	(0.44)
Pathology FE	Yes	Yes	Yes	Yes	-	-	-	-
Radiologist FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	41920	19080	41920	19080	17455	17455	9538	9538

Note: This table summarizes the average treatment effects (ATE) of different information environments on the absolute value of the difference between the radiologist probability and AI probability (column (1) and (2)); absolute value of the difference between the radiologist probability and the diagnostic standard (columns (3) and (4)); and radiologists' effort measured in terms of active time and clicks (columns (5), (6), (7) and (8)). We either pool across all designs (All Designs) or condition on only design 1. Results on effort measure excludes five patient-cases with unaccounted time measure, and observations from design 3 because of learning effects in this set-up. Active time is winsorized to the 95th percentile. The results are for the two top-level pathologies with AI predictions, airspace opacity and cardiomediastinal abnormality. Standard errors are two-way clustered at the radiologist and patient-case level in parenthesis. Robustness by design can be found in section C.3.5.

By Design

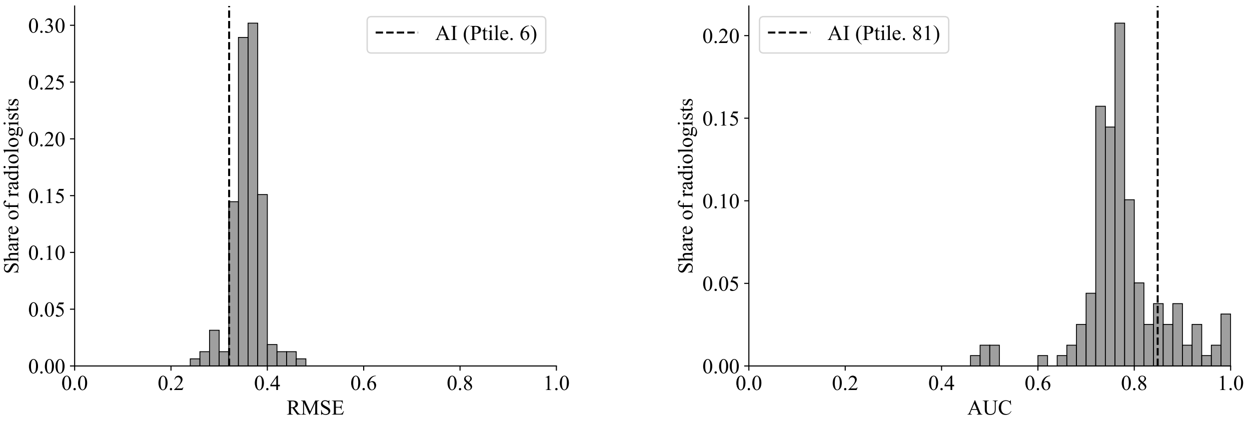
Design 1

This section presents summaries of radiologist accuracy and treatment effect estimates using data from design 1. We do not present treatment effects conditional on radiologist prediction as same cases are not read in the design.

Figure C.12: Comparing AI performance to radiologists

(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

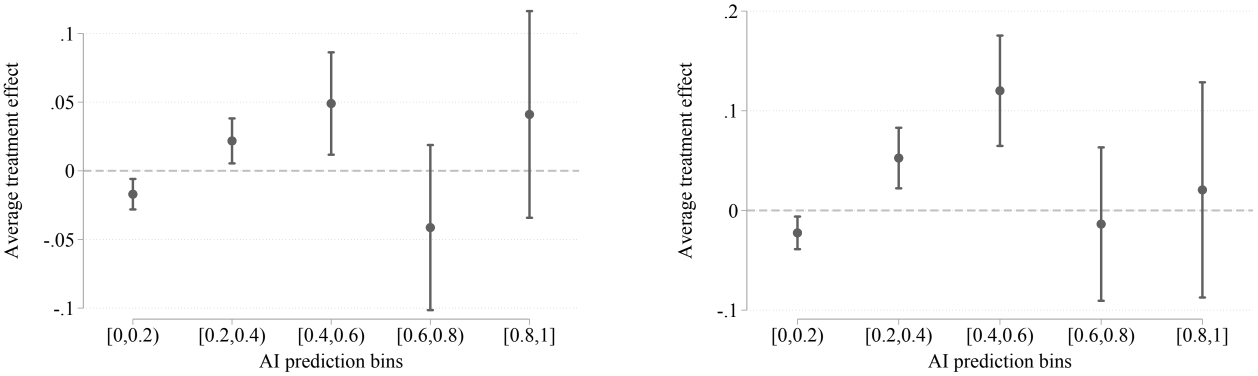


Note: Main specifications similar to figure 3.1 with the exception that observations are from design 1 only.

Figure C.13: Conditional treatment effect given AI prediction

(a) Deviation from diagnostic standard

(b) Incorrect decision



Note: Main specifications similar to figure 3.4 with the exception that observations are from design 1 only.

Design 2

This section presents summaries of key variables, radiologist accuracy and treatment effect estimates using data from design 2.

Table C.10: Summary statistics

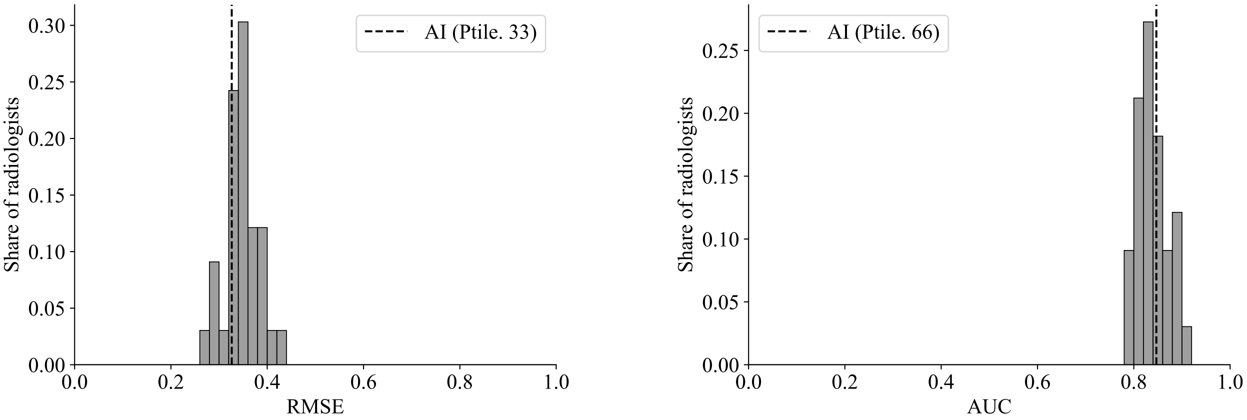
	Mean	SD
	(1)	(2)
Reported Probability	0.245	0.278
Decision	0.400	0.490
Deviation from Diagnostic Standard	0.232	0.265
Deviation from AI	0.172	0.159
Correct Decision	0.620	0.485
Active time	165.6	115.8
Observations	15, 840	
Radiologists	33	

Note: This table presents summary statistics of design 2 similar to table 3.1.

Figure C.14: Comparing AI performance to radiologists

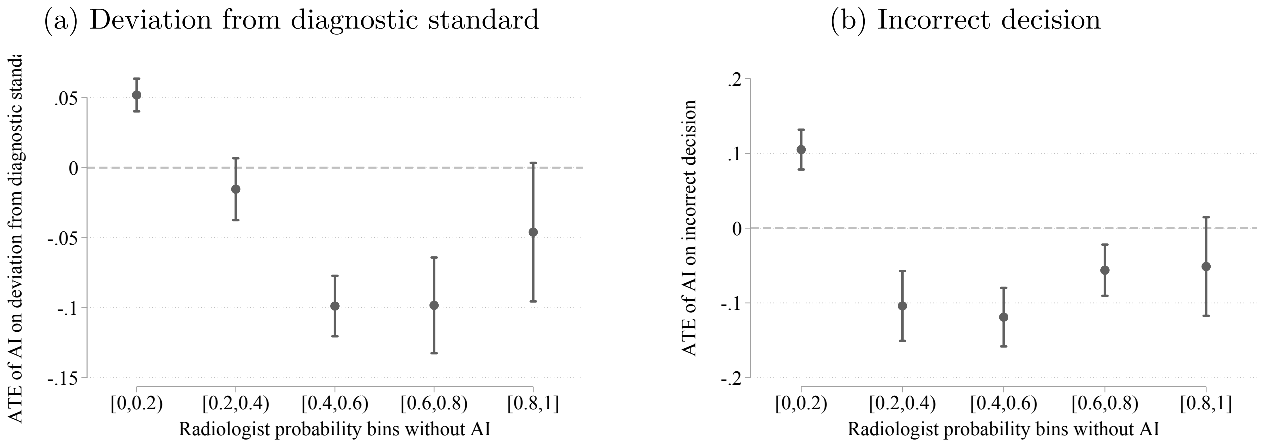
(a) RMSE radiologists and AI

(b) AUROC radiologists and AI



Note: Main specifications similar to figure 3.1 with the exception that observations are from design 2 only.

Figure C.15: Conditional treatment effect given radiologist prediction

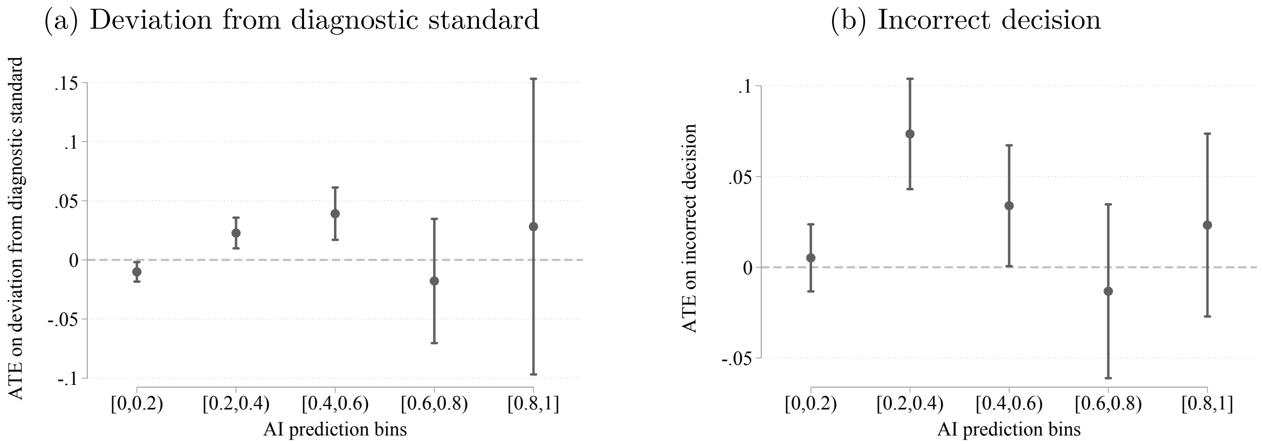


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.3 with the exception that observations are from design 2 only.

Figure C.16: Conditional treatment effect given AI prediction



(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with the exception that observations are from design 2 only.

Table C.11: Average treatment effects

	Deviation from AI	**Deviation from Diagnostic Standard**	**Effort Measures**	
			Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)
AI × CH	−0.001	0.003	−1.61	0.15
	(0.003)	(0.004)	(3.46)	(0.64)
AI	−0.034	0.004	7.14	1.17
	(0.004)	(0.004)	(2.25)	(0.53)
CH	0.001	−0.008	8.08	0.21
	(0.002)	(0.003)	(2.36)	(0.44)
Control Mean	0.189	0.234	154.14	47.18
	(0.009)	(0.013)	(5.72)	(1.75)
Pathology FE	Yes	Yes	-	-
Observations	15840	15840	7917	7917

Note: Main specifications similar to table C.7 with the exception that observations are from design 2 only.

Design 3

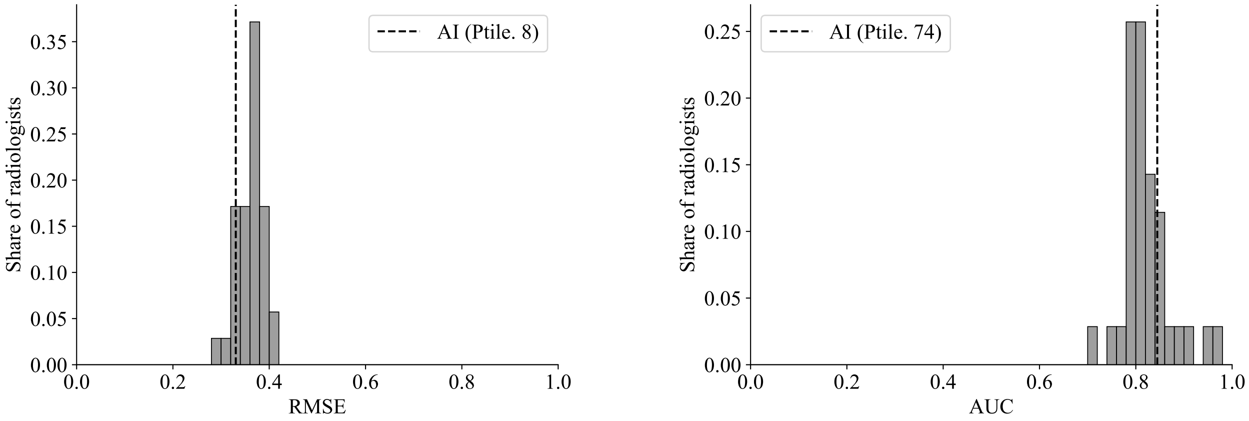
This section presents summaries of key variables, radiologist accuracy and treatment effect estimates using data from design 3.

Table C.12: Summary statistics

	Mean	SD
	(1)	(2)
Reported Probability	0.240	0.322
Decision	0.231	0.421
Deviation from Diagnostic Standard	0.212	0.297
Deviation from AI	0.216	0.182
Correct Decision	0.785	0.411
Active time	154.8	168.0
Observations	7,000	
Radiologists	35	

Note: This table presents summary statistics of design 3 similar to table 3.1.

Figure C.17: Comparing AI performance to radiologists

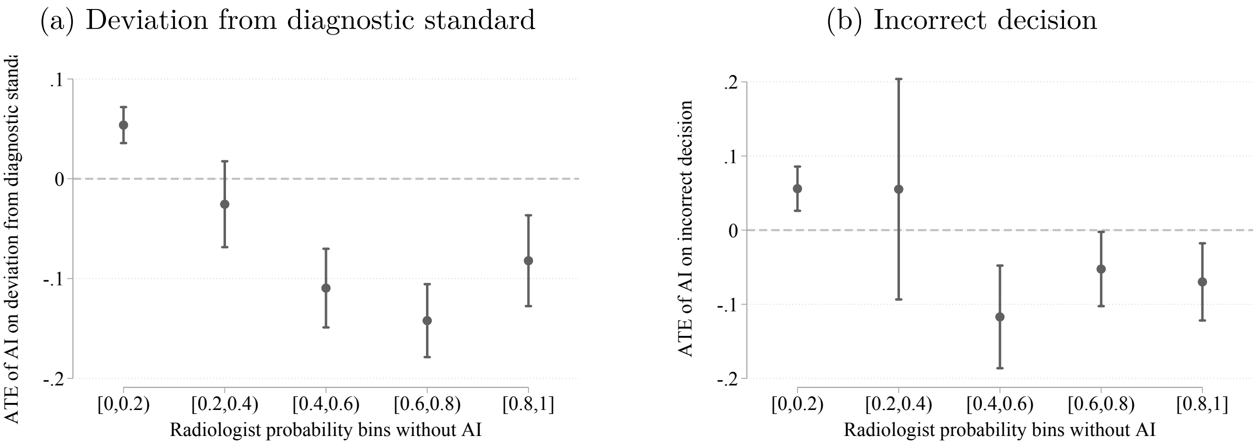


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

Note: Main specifications similar to figure 3.1 with the exception that observations are from design 3 only.

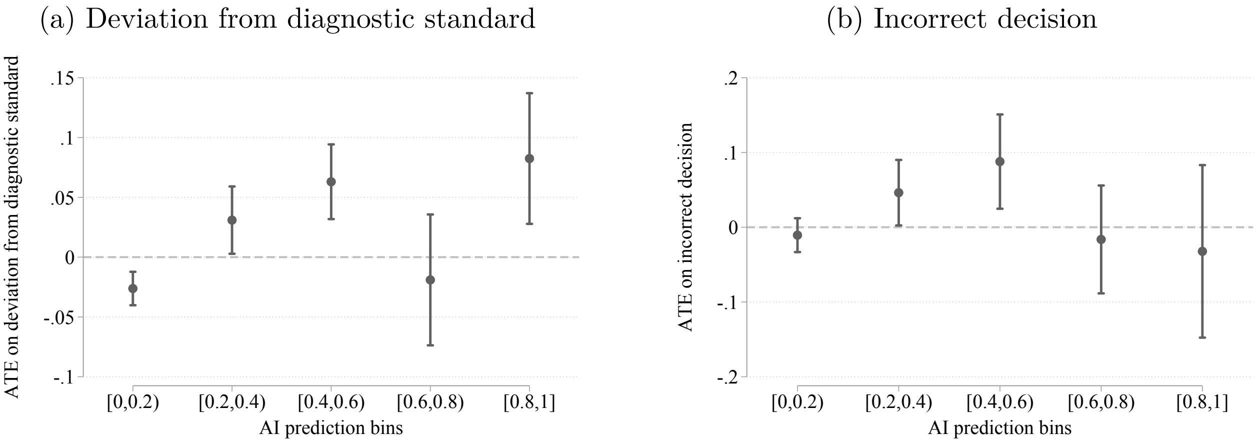
Figure C.18: Conditional treatment effect given radiologist prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.3 with the exception that observations are from design 3 only.

Figure C.19: Conditional treatment effect given AI prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with the exception that observations are from design 3 only.

Table C.13: Average treatment effects

	Deviation from AI	**Deviation from Diagnostic Standard**
Treatment	(1)	(2)
AI × CH	0.009	0.010
	(0.008)	(0.012)
AI	−0.050	−0.003
	(0.010)	(0.008)
CH	−0.003	−0.010
	(0.009)	(0.013)
Control Mean	0.241	0.216
	(0.011)	(0.015)
Pathology FE	Yes	Yes
Observations	7000	7000

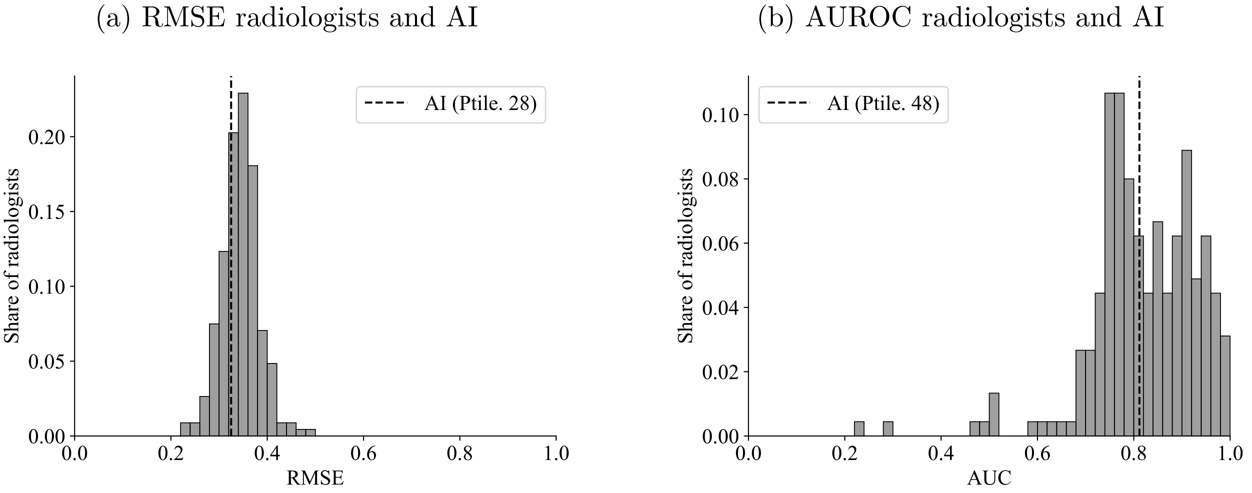
Note: Main specifications similar to table C.7 with the exception that observations are from design 3 only.

By Definition of Diagnostic Standards

Experiment leave-one-out Diagnostic Standard

This section computes the main results using a diagnostic standard constructed using a leave-one-out average of assessments by radiologists participating in the experiment in the treatment arm with clinical history but no AI assistance. Specifically, for each radiologist r and patient case i we construct $\omega_{ir} = 1\left[\sum_{r' \neq r} \frac{\pi(\omega_i = 1 | s_{ir}^E)}{N_i - 1} > 0.5\right]$.

Figure C.20: Comparing AI performance to radiologists

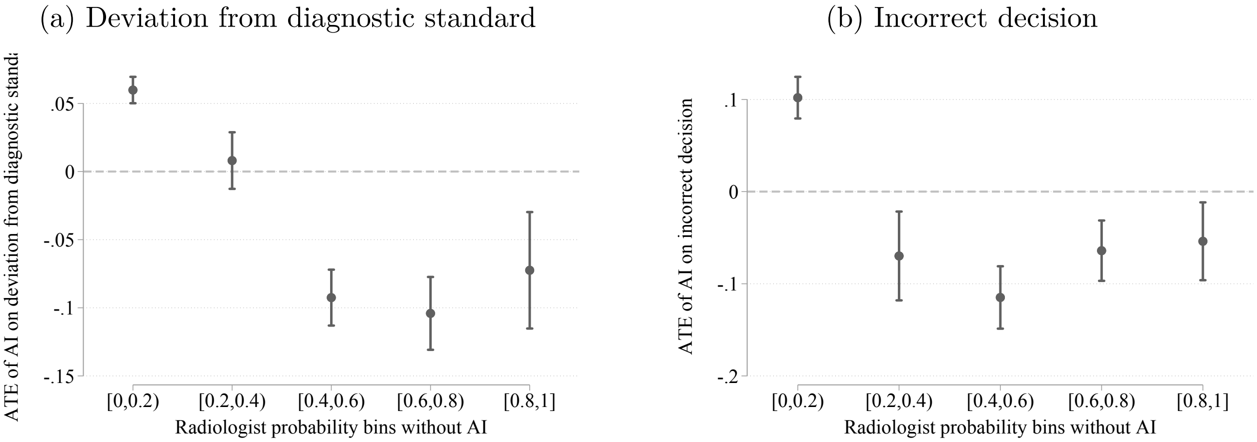


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

Note: Main specifications similar to figure 3.1 with the exception that diagnostic standard is constructed using experiment leave-one-out average.

Figure C.21: Conditional treatment effect given radiologist prediction

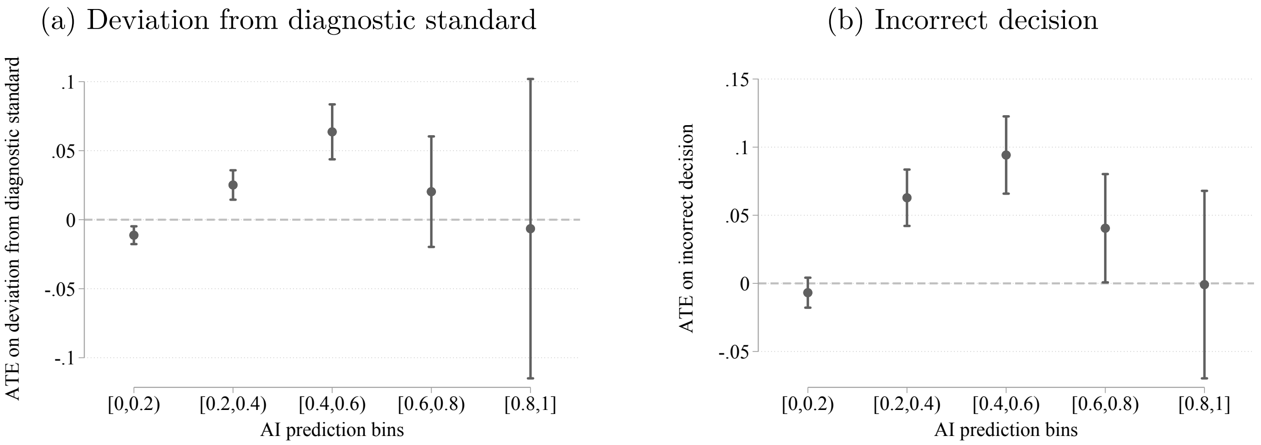


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.3 with the exception that the diagnostic standard is constructed using experiment leave-one-out average.

Figure C.22: Conditional treatment effect given AI prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with the exception that the diagnostic standard is constructed using experiment leave-one-out average.

Table C.14: Average treatment effects

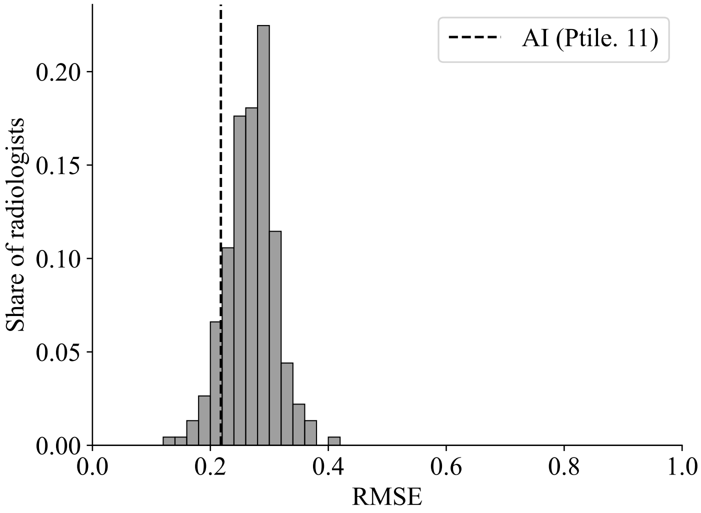
	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	0.004 (0.005)
AI	0.009 (0.004)
CH	−0.012 (0.004)
Control Mean	0.220 (0.010)
Pathology FE	Yes
Observations	41920

Note: Main specifications similar to table C.7 with the exception that the diagnostic standard is constructed using experiment leave-one-out average.

Continuous Diagnostic Standard

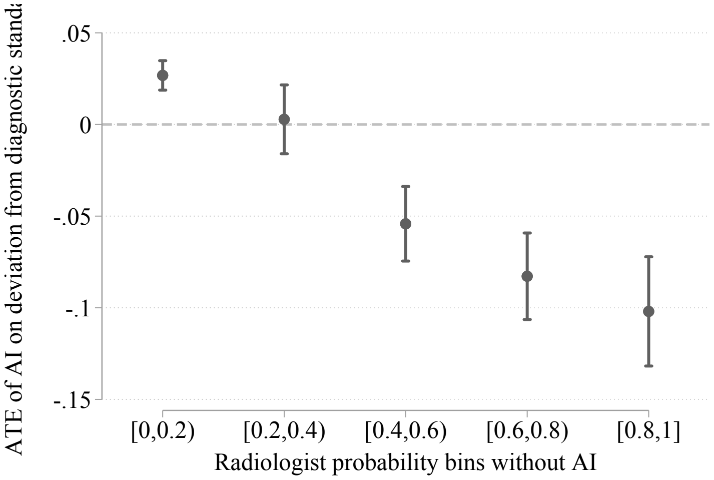
This section computes the main results using a continuous diagnostic standard constructed using a simple average of the diagnostic standard labelers' probability assesment.

Figure C.23: RMSE radiologists and AI



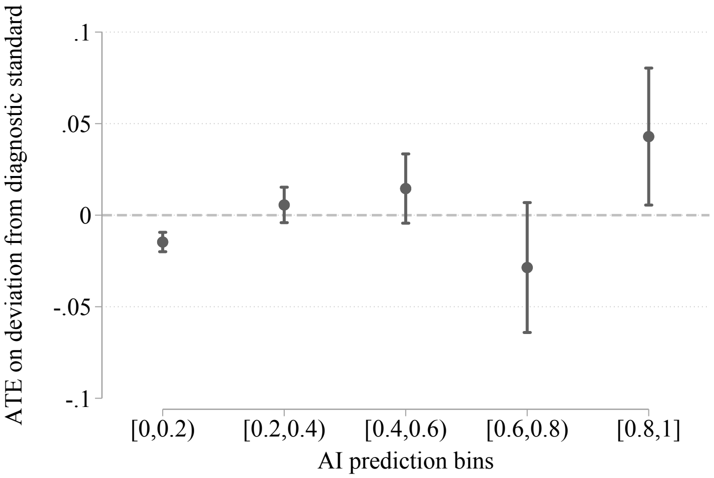
Note: Main specifications similar to figure 3.1 with the exception that the diagnostic standard is constructed using continuous values.

Figure C.24: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 3.3 with the exception that the diagnostic standard is constructed using continuous values.

Figure C.25: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3.4 with the exception that the diagnostic standard is constructed using continuous values.

Table C.15: Average treatment effects

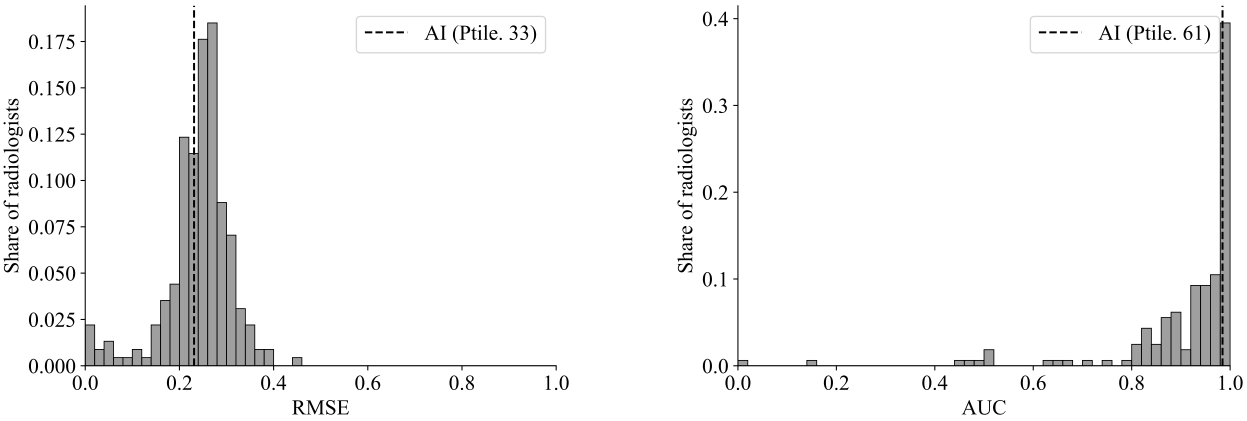
	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	0.002 (0.004)
AI	−0.006 (0.003)
CH	−0.007 (0.003)
Control Mean	0.183 (0.007)
Pathology FE	Yes
Observations	41920

Note: Main specifications similar to table C.7 with the exception thatthe diagnostic standard is constructed using continuous values.

Excluding Cases where the Diagnostic Standard is Uncertain

This section computes the main results using only cases where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level.

Figure C.26: Comparing AI performance to radiologists

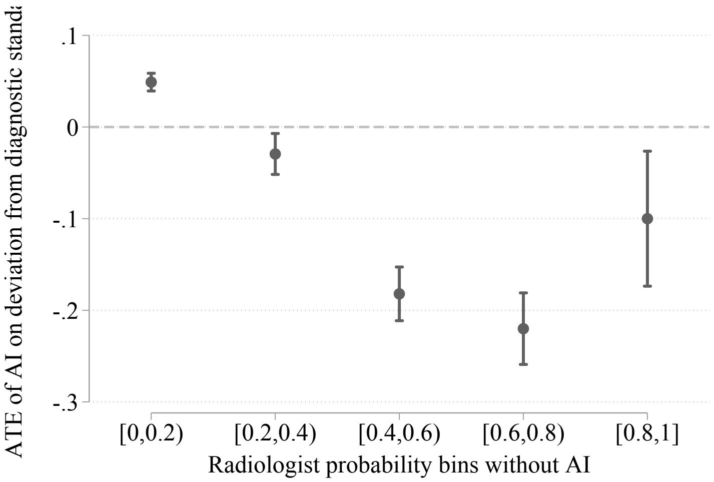


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

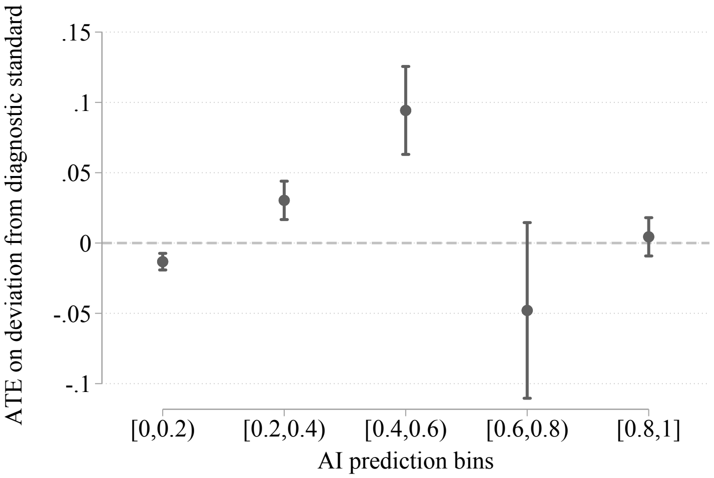
Note: Main specifications similar to figure 3.1 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

Figure C.27: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 3.1 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

Figure C.28: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3.4 3.3 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

Table C.16: Average treatment effects

	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	−0.004 (0.005)
AI	0.007 (0.004)
CH	−0.004 (0.004)
Control Mean	0.138 (0.008)
Pathology FE	Yes
Observations	27703

ote: Main specifications similar to table C.7 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

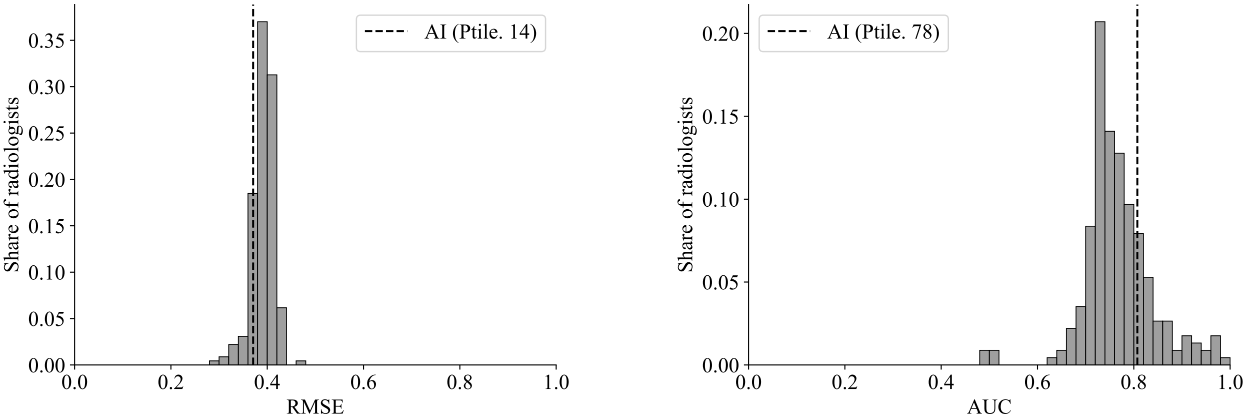
Conservative Diagnostic Standard

This section computes the main results using a binary diagnostic standard with a lower, more conservative cutoff of 0.3 instead of 0.5. That is, $\omega_i = 1\left[\sum_r \pi_r \left(\omega_i = 1 | s_{i,r}^E\right)/5 > 0.3\right]$

Figure C.29: Comparing AI performance to radiologists

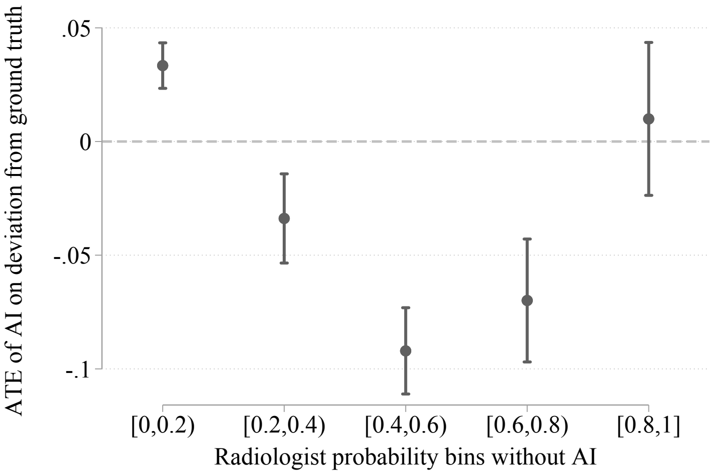
(a) RMSE radiologists and AI

(b) AUROC radiologists and AI



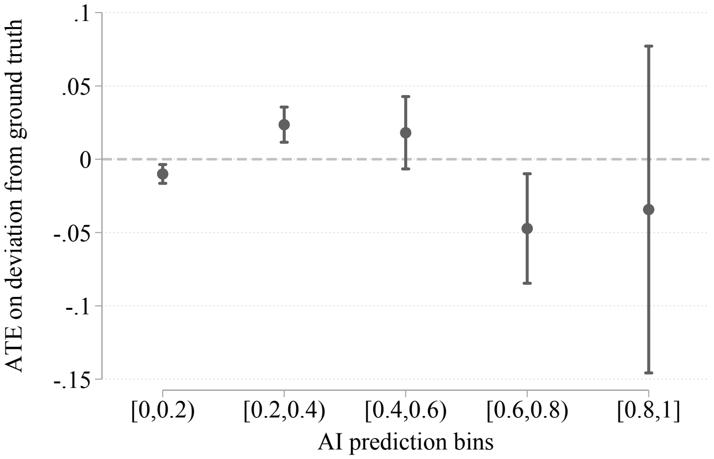
Note: Main specifications similar to figure 3.1 with the exception that binary diagnostic standard uses a lower cutoff at 0.3.

Figure C.30: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 3.3 with the exception that the binary diagnostic standard uses a lower cutoff at 0.3.

Figure C.31: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3.4 with the exception that the binary diagnostic standard uses a lower cutoff at 0.3.

Table C.17: Average treatment effects

	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	0.004
	(0.005)
AI	−0.001
	(0.004)
CH	−0.012
	(0.005)
Control Mean	0.248
	(0.011)
Pathology FE	Yes
Observations	36280

Note: Main specifications similar to table C.7 with the exception that the binary diagnostic standard uses a lower cutoff at 0.3.

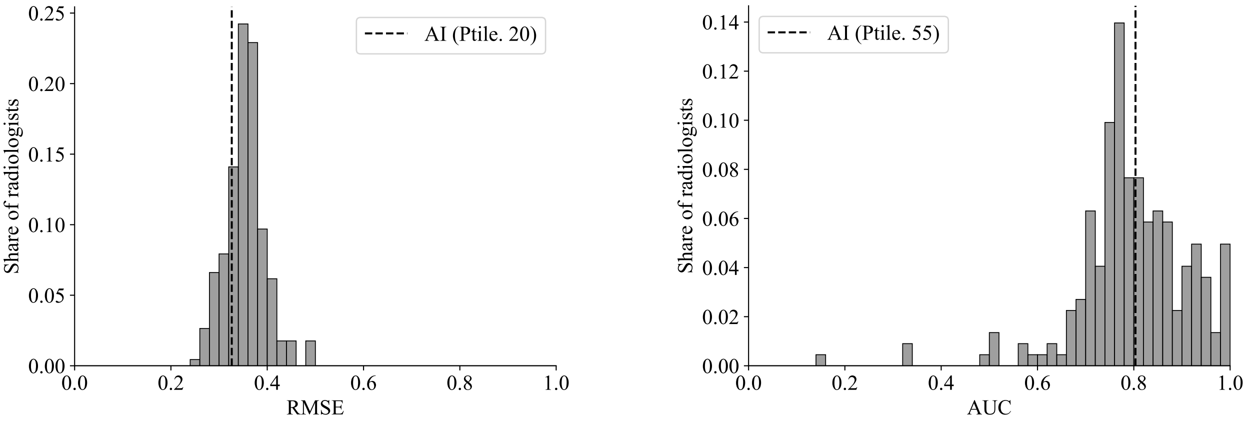
Internally Constructed Diagnostic Standard Excluding Cases with AI and Clinical History

This section computes the main results using a internally constructed diagnostic standard which excludes cases where the radiologist received AI support or clinical history.

Figure C.32: Comparing AI performance to radiologists

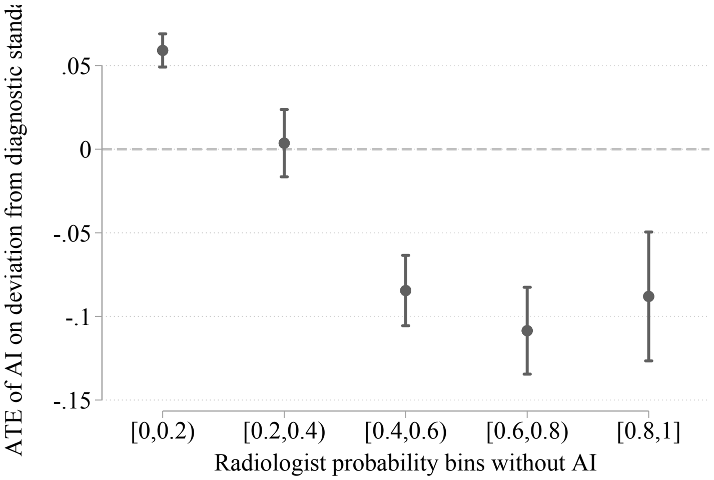
(a) RMSE radiologists and AI

(b) AUROC radiologists and AI



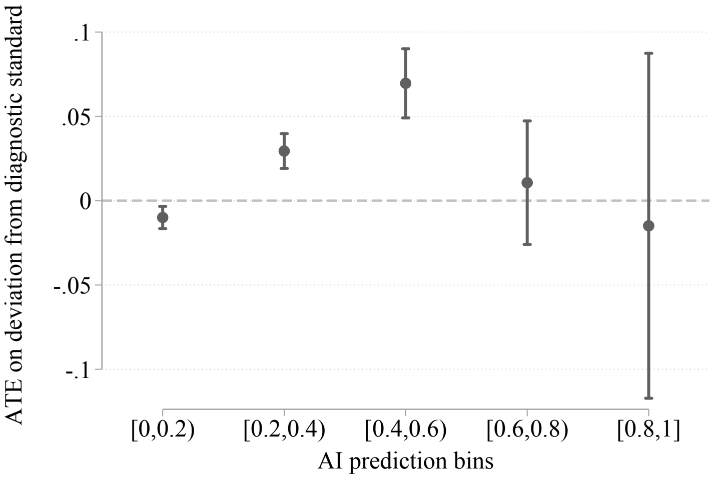
Note: Main specifications similar to figure 3.1 with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

Figure C.33: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 3.3 with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

Figure C.34: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3.4with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

Table C.18: Average treatment effects

	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	−0.007
	(0.005)
AI	0.016
	(0.004)
CH	0.003
	(0.004)
Control Mean	0.214
	(0.009)
Pathology FE	Yes
Observations	41920

Note: Main specifications similar to table C.7 with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

Testing for Order Effects

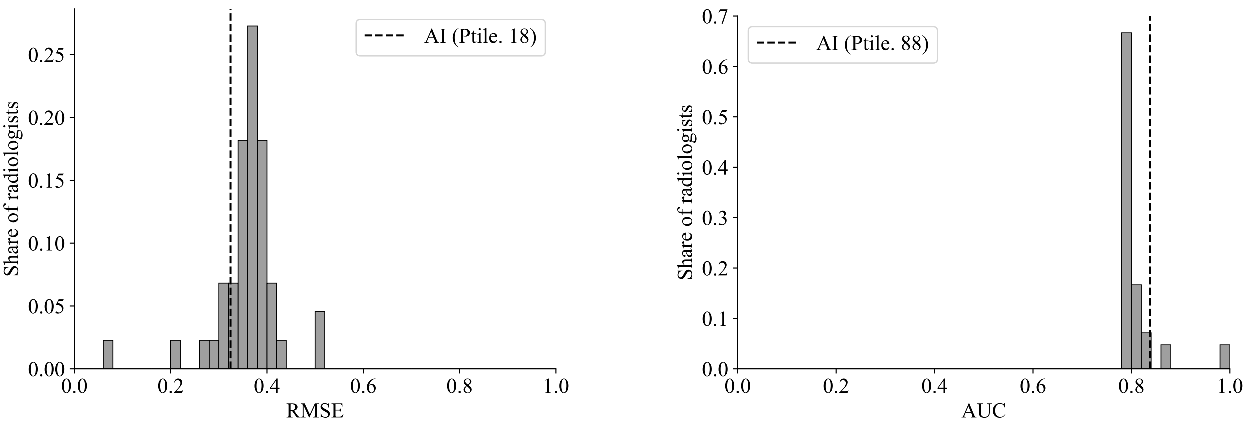
First Treatment (only Design 1 and Design 2)

The following graphs contain only those cases from the treatment group that the subjects encountered first. This includes the first 15 reads from design 1 and the first 5 reads from design 2. This exercise is to check if the treatment effects for all the reads is different than for the first reads.

Figure C.35: Comparing AI performance to radiologists

(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

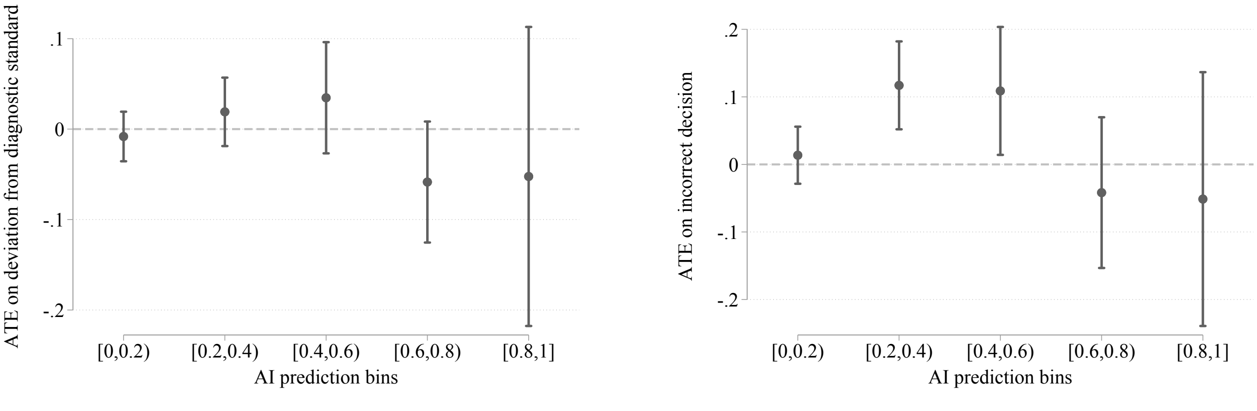


Note: Main specifications similar to figure 3.1 with the exception that observations are from the first treatment received in designs 1 and 2 only.

Figure C.36: Conditional treatment effect given AI prediction

(a) Deviation from diagnostic standard

(b) Incorrect decision



Note: Main specifications similar to figure 3.4 with the exception that observations are from the first treatment received in designs 1 and 2 only.

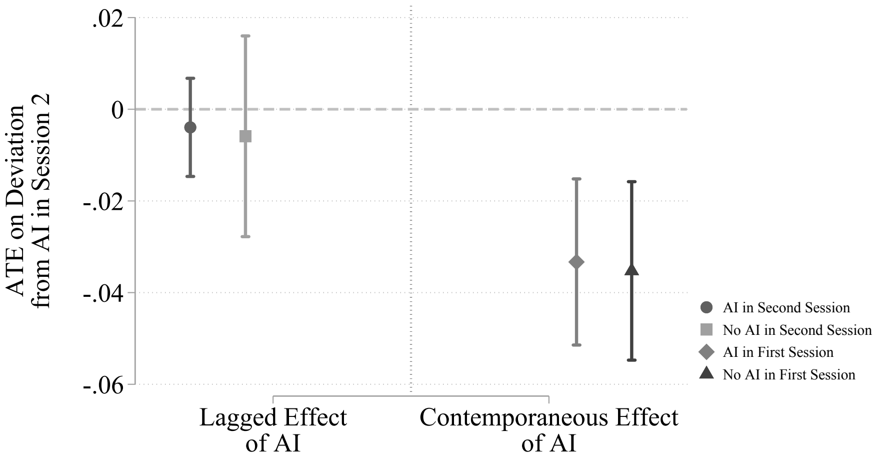
Table C.19: Average treatment effects

	Deviation from AI	**Deviation from Diagnostic Standard**	**Effort Measures** Active Time	 Clicks
Treatment	(1)	(2)	(3)	(4)
AI × CH	−0.020	−0.015	6.29	0.97
	(0.018)	(0.024)	(24.80)	(5.56)
AI	−0.040	0.011	−6.34	1.48
	(0.013)	(0.018	(16.06)	(3.77)
CH	−0.014	−0.004	25.73	2.04
	(0.013)	(0.017)	(19.25)	(4.27)
Control Mean	0.235	0.231	180.48	42.49
	(0.009)	(0.015)	(13.39)	(2.76)
Pathology FE	Yes	Yes	-	-
Observations	5100	5100	2550	2550

Note: Main specifications similar to table C.7 with the exception that observations are from the first treatment received in designs 1 and 2 only.

Previous Exposure to AI (Design 2)

Figure C.37: Conditional treatment effect given AI prediction



Note: This graph shows the treatment effects of the deviation from AI in the second session because of receiving AI signal in the first session conditional on receiving AI signal in the second session (Lagged Effect). On the other hand, the contemporaneous effects show the deviation from AI in the second session given the participants receive AI in the second session, conditional on receiving AI signal in the second session. These graphs are valid only for design 2 as participants see the same image but in a different information environment.

Incentives

This section tests if incentives for assessment accuracy promote radiologists to make more accurate assessments. We find that the incentives do not play a significant role in getting a correct response. The effect of incentives are estimated using the following regression specification and the results are shown in Table C.20.

$$\begin{aligned}
Y_{irt} =& \gamma_{h_i} + \gamma_{INC} \cdot d_{INC}(r) \\
& + \gamma_{CH} \cdot d_{CH}(t) + \gamma_{CH \times INC} \cdot d_{CH}(t).d_{INC}(r) \\
& + \gamma_{AI} \cdot d_{AI}(t) + \gamma_{AI \times INC} \cdot d_{AI}(t).d_{INC}(r) \\
& + \gamma_{AI \times CH} \cdot d_{CH}(t) \cdot d_{AI}(t) + \gamma_{AI \times CH \times INC} \cdot d_{CH}(t) \cdot d_{AI}(t).d_{INC}(r) + \varepsilon_{irt}
\end{aligned}$$

where Y_{irt} is an outcome variable of interest for radiologist r diagnosing patient case-pathology i and treatment t, and γ_{h_i} are pathology fixed effects. Here CH refers to cases with access to clinical history information, AI to cases with AI predictions and INC refers to incentivized cases.

Table C.20: Effect of incentives

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**	
	Top-Level with AI	**Pooled with AI**	**Top-Level with AI**	**Pooled with AI**	**Active Time**	**Clicks**
Treatment	(1)	(2)	(3)	(4)	(5)	(6)
AI × CH	−0.001	−0.003	−0.002	−0.002	−9.37	−1.50
	(0.006)	(0.002)	(0.012)	(0.004)	(7.45)	(1.73)
AI	−0.033	−0.013	0.003	0.001	11.26	2.44
	(0.006)	(0.003)	(0.009)	(0.003)	(5.54)	(1.32)
CH	−0.000	0.001	−0.012	−0.004	11.10	0.87
	(0.005)	(0.002)	(0.008)	(0.003)	(5.07)	(1.28)
Control Mean	0.223	0.112	0.221	0.083	156.22	39.27
	(0.008)	(0.003)	(0.012)	(0.005)	(8.80)	(1.94)
AI × CH × Incentives	0.010	0.006	0.004	0.002	17.08	3.04
	(0.009)	(0.003)	(0.016)	(0.006)	(11.83)	(2.57)
AI × Incentives	−0.021	−0.007	−0.002	0.001	−12.72	−2.37
	(0.008)	(0.003)	(0.012)	(0.004)	(8.06)	(1.78)
CH × Incentives	−0.006	−0.003	0.004	0.003	−5.95	−1.22
	(0.007)	(0.003)	(0.013)	(0.005)	(8.30)	(1.77)
Control Mean × Incentives	0.006	0.002	−0.000	−0.003	−3.53	−0.77
	(0.009)	(0.003)	(0.011)	(0.004)	(12.10)	(2.72)
Pathology FE	Yes	Yes	Yes	Yes	-	-
Observations	26080	169520	26080	169520	9538	9538
F-stat	1.61	1.35	.18	.59	1.08	.71
P>F-stat	.17	.25	.95	.67	.37	.58

Note: This table summarizes the average treatment effects (ATE) of different information environments on the (1) absolute value of the difference between the radiologist probability and AI algorithm probability (Columns (1) and (2)), absolute value of the difference between the radiologist probability and the diagnostic standard (Columns (3) and (4)) and radiologists' effort measured in terms of active time and clicks for all and the second-half images (Columns (5) and (6)). The F-statistic tests for the joint significance of the four incentivized groups. Top-level specification includes two pathologies: airspace opacity and cardiomediastinal abnormality while Pooled AI includes all the pathologies with AI predictions excluding abnormality and support device hardware. Only cases in design 1 and design 3 are considered. Two-way clustered standard errors at the radiologist and patient-case level are in parenthesis.

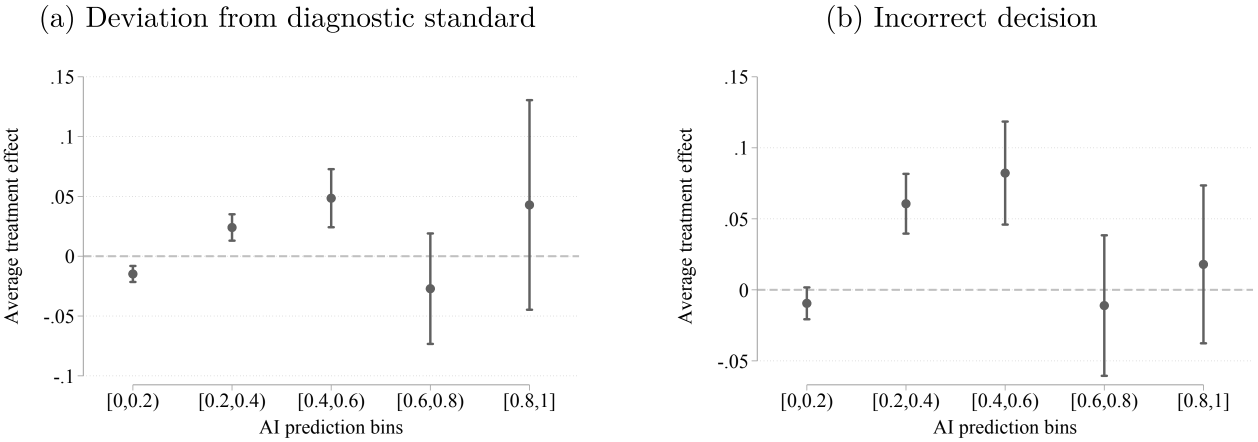
Controlling for sequence number and session

Figure C.38 uses the following specification that controls for the sequence number in which the participants saw a particular case within one experiment session and the session dummies for the different designs and experiment sessions to estimate the heterogeneous treatment effects. There are four sessions in Design 2, whereas Design 1 and 3 have only one session.

$$Y_{irt} \;=\; \gamma_{h_i} + \gamma_{AI} \cdot d_{AI}(t) + \sum_{g} \Big[\gamma_g \cdot d_g(s_i^A) + \gamma_{AI \times g} \cdot d_{AI}(t).d_g(s_i^A) \Big] + \gamma_{w_{irt}} + \gamma_{m_{irt}} + \varepsilon_{irt}$$

where Y_{irt} is an outcome variable of interest for radiologist r diagnosing patient case-pathology i and treatment t, γ_{h_i} are pathology fixed effects, $\gamma_{w_{irt}}$ are sequence number dummies and $\gamma_{m_{irt}}$ are session dummies. Here, g is defined as an index for an interval of the AI signal range where $1 \leq g \leq 5$. A case is said to be in an interval g conditional on the signal value for the given patient-case.

Figure C.38: Conditional treatment effect given AI prediction



(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with additional controls for rounds and session.

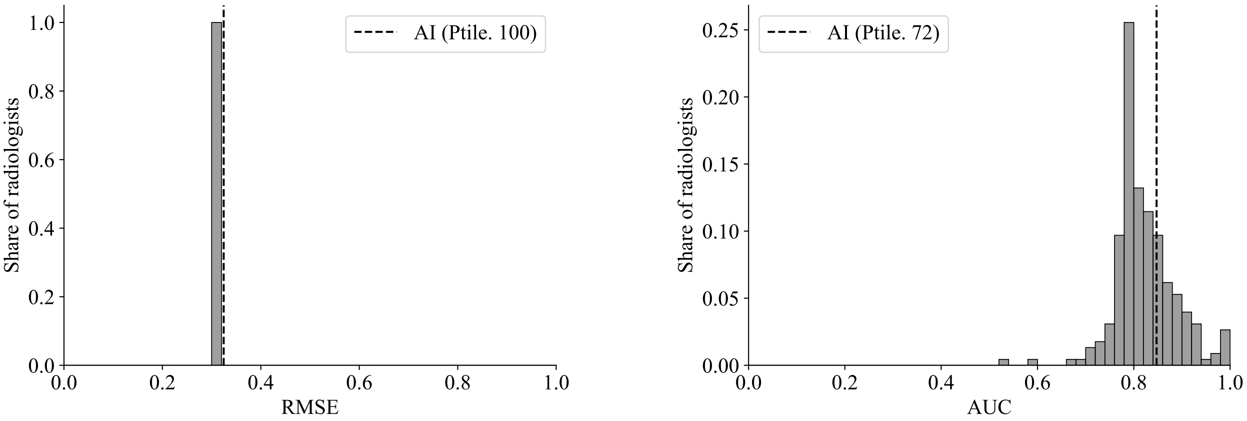
Table C.21: Average treatment effects

	Deviation	**Deviation**	**Effort Measures**	
	from AI	**from Diagnostic Standard**	**Active Time**	**Clicks**
Sessions	(1)	(2)	(3)	(4)
Design 2: Session 1	−0.018	0.027	47.44	16.34
	(0.010)	(0.013)	(10.61)	(2.56)
Design 2: Session 2	−0.035	0.012	2.54	8.33
	(0.010)	(0.011)	(9.28)	(2.36)
Design 2: Session 3	−0.033	0.002	−18.90	5.37
	(0.010)	(0.011)	(9.13)	(2.46)
Design 2: Session 4	−0.037	0.004	−32.53	3.06
	(0.009)	(0.011)	(7.07)	(2.25)
Design 3	0.019	−0.008	-	-
	(0.009)	(0.011)	-	-
Control Mean	0.220	0.218	158.54	39.01
	(0.006)	(0.010)	(5.78)	(1.33)
Design 2: Session 1 × AI	0.001	0.007	6.87	0.89
	(0.006)	(0.009)	(5.80)	(1.25)
Design 2: Session 2 × AI	0.007	−0.004	0.11	−0.50
	(0.007)	(0.009)	(5.32)	(1.21)
Design 2: Session 3 × AI	0.002	0.003	0.70	−0.61
	(0.008)	(0.009)	(4.54)	(1.25)
Design 2: Session 4 × AI	0.010	0.005	−0.21	0.12
	(0.007)	(0.011)	(4.59)	(1.15)
Design 3 × AI	−0.006	−0.000	-	-
	(0.009)	(0.008)	-	-
Control Mean × AI	−0.040	0.002	4.49	1.27
	(0.004)	(0.005)	(3.08)	(0.66)
Pathology FE	Yes	Yes	-	-
Observations	41920	41920	17455	17455
F-stat	.71	.29	.48	.33
P>F-stat	.62	.92	.75	.86

Note: Main specifications similar to table C.7 with additional control variables for sequence number of a particular case and the experiment session. Due to the high volume of sequence numbers, we do not show them in this table but account for them. Design 1 session dummy is ommitted due to collinearity and is thus the control mean.

Calibrated radiologist probability

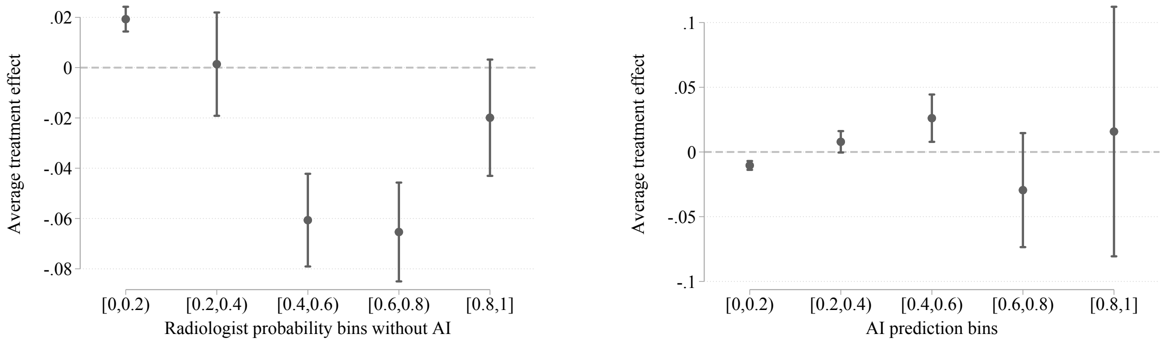
Figure C.39: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

Note: Main specifications similar to figure 3.1 with the exception that reported probability of the radiologists is calibrated to the diagnostic standard.

Figure C.40: Deviation from diagnostic standard


(a) Conditional on radiologist signal

(b) Conditional on AI signal

Note: Main specifications similar to figure 3.3 and figure 3.4 with the exception that reported probability of the radiologists is calibrated to the diagnostic standard and there are no separate results for ATE on incorrect decision.

Table C.22: Average treatment effects

	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	−0.001 (0.004)
AI	−0.000 (0.003)
CH	−0.002 (0.003)
Control Mean	0.187 (0.010)
Pathology FE	Yes
Observations	41917

Note: Main specifications similar to table C.7 with the exception that reported probability of the radiologists is calibrated to the diagnostic standard and hence only the ATE on deviation from the diagnostic standard is reported.

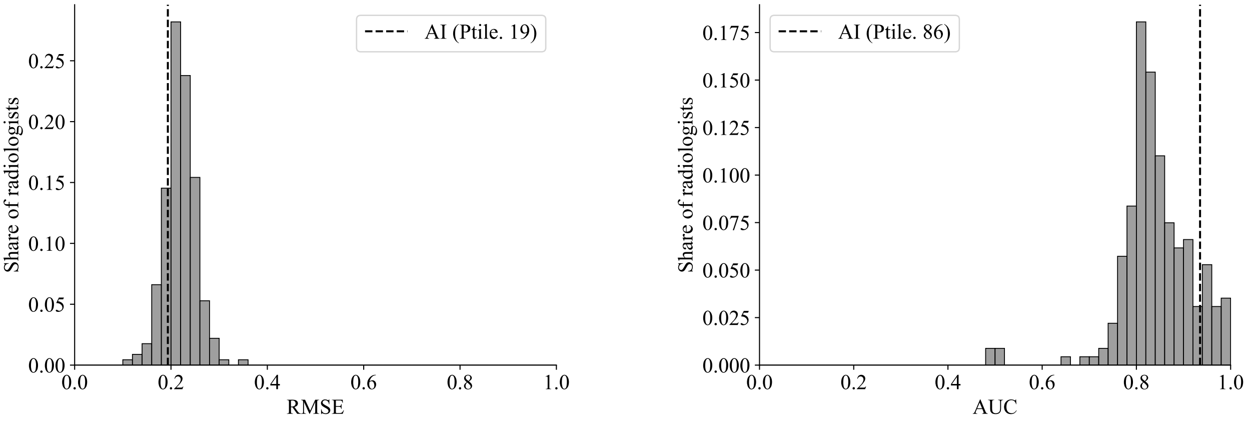
All Pathologies and Abnormal with AI

All Pathologies with AI

Figure C.41: Comparing AI performance to radiologists

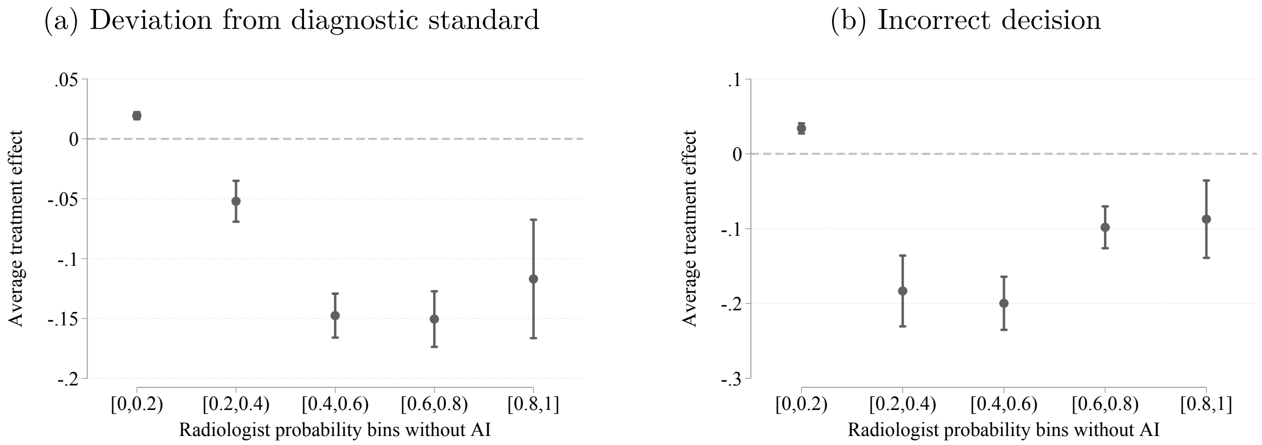
(a) RMSE radiologists and AI

(b) AUROC radiologists and AI



Note: Main specifications similar to figure 3.1 with the exception that all patholgies with AI are considered.

Figure C.42: Conditional treatment effect given radiologist prediction

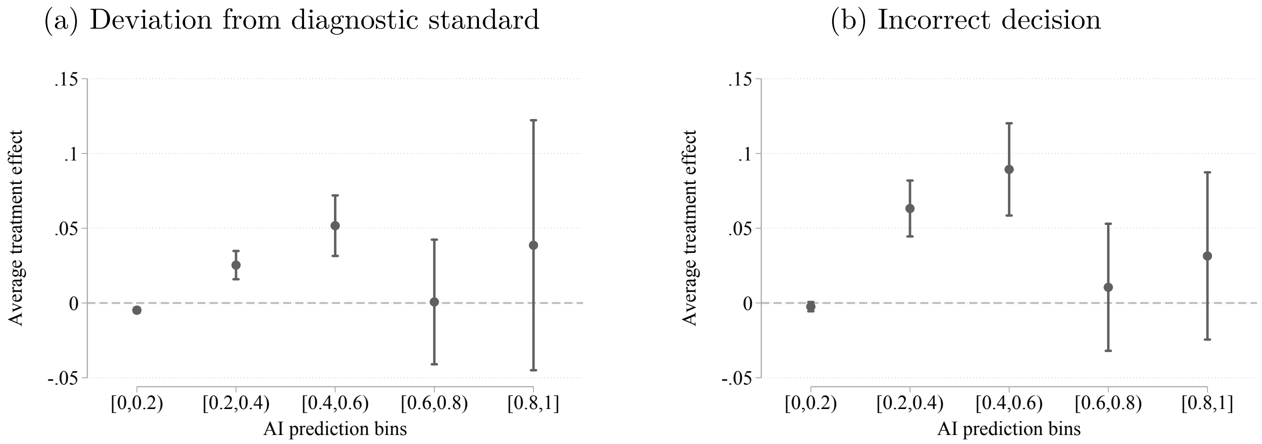


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with the exception that all patholgies with AI are considered.

Figure C.43: Conditional treatment effect given AI prediction



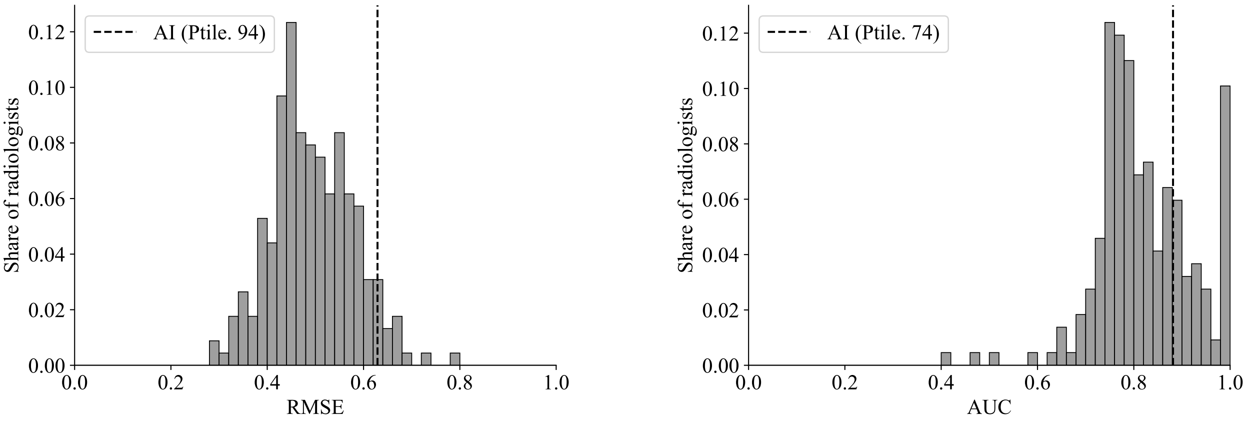
(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with the exception that all patholgies with AI are considered.

Abnormal

Figure C.44: Comparing AI performance to radiologists

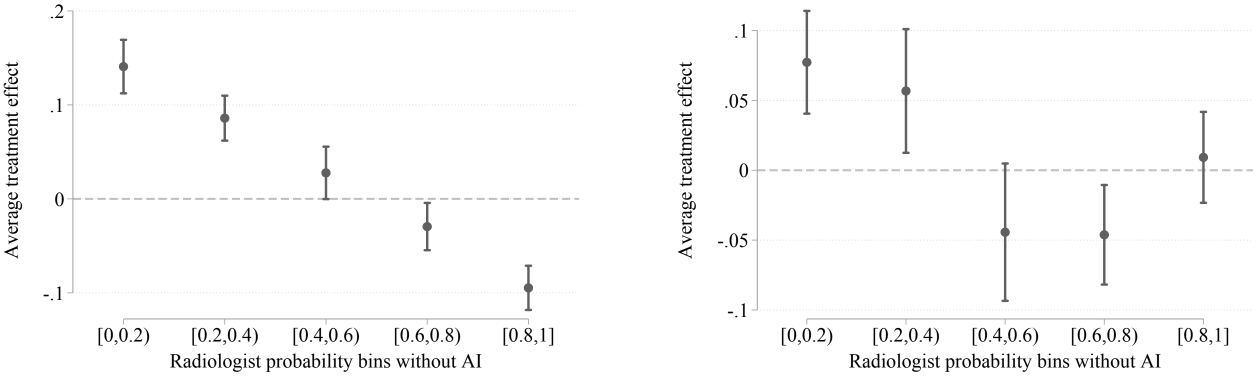


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

Note: Main specifications similar to figure 3.1 with the exception that only the abnormal pathology is considered.

Figure C.45: Conditional treatment effect given radiologist prediction

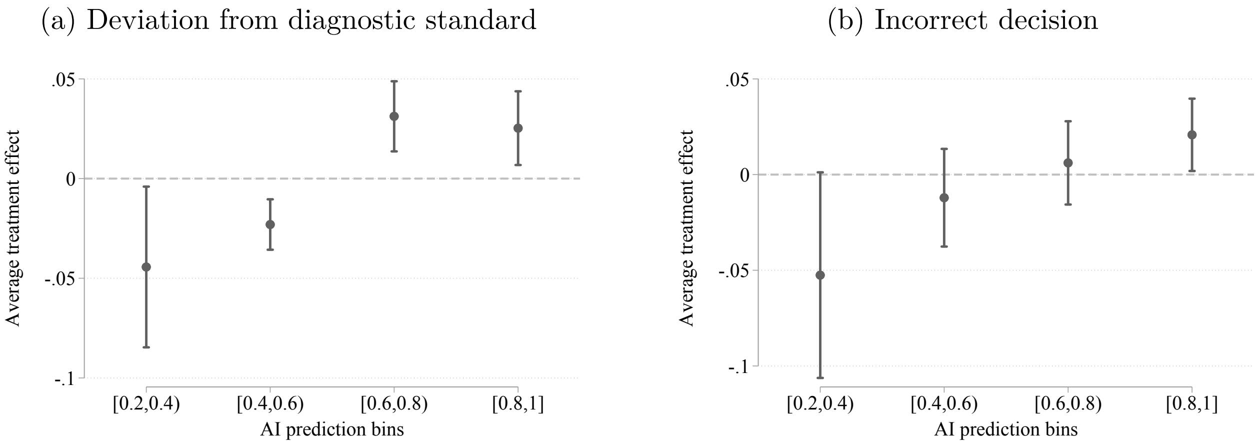


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with the exception that only the abnormal pathology is considered.

Figure C.46: Conditional treatment effect given AI prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 3.4 with the exception that only the abnormal pathology is considered.

Table C.23: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		
	Pooled with AI	Abnormal	Pooled	Pooled with AI	Abnormal
Treatment	(1)	(2)	(3)	(4)	(5)
AI × CH	0.000	0.010	0.000	−0.000	0.002
	(0.001)	(0.005)	(0.001)	(0.002)	(0.008)
AI	−0.016	−0.062	−0.001	0.001	0.013
	(0.001)	(0.005)	(0.001)	(0.001)	(0.007)
CH	−0.000	−0.003	−0.001	−0.002	−0.005
	(0.001)	(0.004)	(0.001)	(0.001)	(0.006)
Control Mean	0.109	0.279	0.032	0.085	0.419
	(0.003)	(0.009)	(0.002)	(0.004)	(0.012)
Pathology FE	Yes	No	Yes	Yes	No
Observations	272480	20960	2137920	272480	20960

Note: Main specifications similar to table C.7 with the exception that only the abnormal pathology is considered.

C.3.6 Automation Bias Appendix

Conditional Independence

Table C.24 presents evidence that human and AI signals are not conditionally independent. To test the hypothesis of conditional independence, we regress the human report in the treatment arm without AI assistance on the diagnostic standard, the AI score, and the interaction between the diagnostic standard and the AI score. If the signals were conditionally independent, we would observe the AI score to offer no predictive power on the human report after conditioning on the diagnostic standard. As shown in Table C.24 we can reject this null hypothesis.

Table C.24: Test of conditionally independent signals

	Top Level with AI	Pooled with AI	Abnormal
Diagnostic Standard	0.318	0.265	-0.049
	(0.042)	(0.033)	(0.078)
AI Score	0.536	0.620	0.815
	(0.046)	(0.035)	(0.052)
Diagnostic Standard × AI	-0.244	-0.217	0.205
	(0.082)	(0.066)	(0.090)
Constant	0.089	0.044	-0.058
	(0.011)	(0.003)	(0.039)
Observations	11420	57100	5710
R-Squared	0.260	0.301	0.316

Note: Estimates of a regression of the human report in the treatment without AI assistance on the diagnostic standard interacted with the AI score. This table uses data from designs 2 and 3. Standard errors are two-way clustered at the radiologist and patient case level.

Estimating Bayesian Update Terms

Here, we describe the method we use to estimate the Bayesian benchmark $\pi(\omega_i = 1|s_{ih}^H, s_i^A)$. This procedure is done separately for each pathology. We train a random forest classifier that predicts the diagnostic standard based on features including the vector of a radiologist's reported probabilities in the non-AI treatment and the vector of AI predictions. Additional features include radiologist identifiers to allow for heterogeneity in radiologists' assessments, an indicator equal to one if the case was read with clinical history, and summaries of the patient clinical history. We estimate this quantity for various parameterizations of s_{ih}^H and s_i^A described in Section 3.5. These are used in the model testing exercise to understand if radiologists account for the joint distribution of signals when forming their posterior beliefs. The hyperparameters of the model are tuned using grouped cross-validation where observations were grouped by patient id to avoid overfitting with five folds. We impose monotonicity constraints on the model to impose that $\pi(\omega_i = 1|s_{ih}^H, s_i^A)$ is monotonically increasing in all probability inputs. When the model includes clinical history, we provide a summarized patient's clinical record with their sex, weight, temperature, pulse, age, and the number of available and flagged labs. We mean impute these variables when the radiologist does not have access to the clinical history and include an indicator equal to one if the radiologist had access to the clinical history as an additional feature. Below we summarize the performance of these models and the relative value of increasing the dimension of s_{ih}^H and s_i^A.

Table C.25: Summary of Bayesian models

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Airspace Opacity	Accuracy	0.88	0.91	0.88	0.89	0.90	0.90	0.89	0.90	0.89	0.89	0.91	0.90
	AUC	0.89	0.96	0.91	0.92	0.96	0.96	0.94	0.96	0.93	0.93	0.96	0.96
Cardiomediastinal Abnorm.	Accuracy	0.90	0.92	0.90	0.91	0.93	0.92	0.92	0.92	0.91	0.91	0.93	0.93
	AUC	0.90	0.96	0.91	0.93	0.96	0.96	0.94	0.96	0.94	0.94	0.96	0.96
Abnormal	Accuracy	0.88	0.91	0.88	0.89	0.91	0.91	0.90	0.91	0.90	0.90	0.92	0.91
	AUC	0.91	0.96	0.91	0.93	0.96	0.96	0.95	0.96	0.94	0.94	0.96	0.96
Focal s_A		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A			✓			✓	✓		✓			✓	✓
Focal s_E				✓	✓	✓	✓			✓	✓	✓	✓
Other s_E					✓		✓				✓		✓
Clinical History s_E								✓	✓	✓	✓	✓	✓

Note: This table summarizes the models used to estimate the Bayesian benchmark. Each column corresponds to a random forest classification tree with varying signal structures. The rows Focal s^A, Other s^A, Focal s^H, Other s^H, and Clinical History s^H indicate what features are included in the tree. Focal s^A corresponds to the focal pathology's AI score, Other s^A corresponds to vector of AI scores for all pathologies, Focal (Other) s^H includes the radiologist's report without AI assistance on the focal pathology (all pathologies), and Clinical History s^H contains summaries of the patient's clinical history when available to the radiologist.

Model Selection on Additional Pathology Groups

Here, we present the model selection results for the remaining pre-registered pathology groups.

Table C.26: Model selection: top level with AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.27	0.12	0.33	0.29	0.12	0.12	0.19	0.12	0.21	0.23	0.12	0.12
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Own information bias (d)	1.11	1.05	1.09	1.08	1.05	1.05	1.07	1.05	1.07	1.08	1.05	1.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.39	0.25	0.39	0.38	0.25	0.25	0.32	0.25	0.32	0.35	0.25	0.25
	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	13.08	7.68	11.63	15.4	7.59	8.85	7.53	6.25	8.55	9.34	6.72	7.72
MMSC-BIC	-29.11	-9.19	-26.34	-18.36	-9.29	-8.03	-13.57	-10.63	-12.55	-11.76	-10.16	-9.16
Selected moments	13	7	12	11	7	7	8	7	8	8	7	7
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420
R-Squared	0.49	0.42	0.48	0.48	0.42	0.42	0.45	0.42	0.45	0.46	0.42	0.42

Note: This table presents results of the model selection exercise as described in Table 3.2, including the full set of models that were included in the selection procedure.

Table C.27: Model selection: pooled with AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.17	0.09	0.19	0.14	0.10	0.10	0.12	0.10	0.14	0.14	0.10	0.10
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Own information bias (d)	1.12	1.10	1.12	1.12	1.10	1.10	1.11	1.10	1.11	1.11	1.10	1.10
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.37	0.31	0.38	0.36	0.31	0.30	0.33	0.31	0.34	0.35	0.31	0.30
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	11.16	6.44	12.05	4.11	3.62	4.53	5.8	5.22	5.32	4.77	4.43	4.68
MMSC-BIC	-14.16	-10.44	-13.26	-12.77	-13.25	-12.35	-11.07	-11.65	-11.56	-12.11	-12.45	-12.2
Selected moments	9	7	9	7	7	7	7	7	7	7	7	7
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100
R-Squared	0.46	0.42	0.44	0.43	0.42	0.42	0.43	0.42	0.43	0.43	0.42	0.42

Note: This table presents results of the model selection exercise as described in Table 3.2, though this table only includes all pathologies with AI assistance.

Table C.28: Model selection: abnormal

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.09	0.06	0.08	0.09	0.06	0.07	0.05	0.07	0.05	0.07	0.06	0.07
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Own information bias (d)	1.07	1.07	1.07	1.07	1.07	1.07	1.07	1.08	1.07	1.07	1.07	1.08
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.36	0.30	0.33	0.37	0.30	0.31	0.26	0.31	0.26	0.32	0.30	0.31
	(0.07)	(0.06)	(0.07)	(0.08)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	11.99	11.02	12.87	11.1	11.02	10.95	14.45	11.03	14.48	12.64	11.04	11.05
MMSC-BIC	-25.98	-26.95	-25.1	-26.88	-26.95	-27.03	-23.52	-26.94	-23.49	-25.33	-26.93	-26.92
Selected moments	12	12	12	12	12	12	12	12	12	12	12	12
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710
R-Squared	0.35	0.32	0.32	0.34	0.32	0.32	0.30	0.32	0.30	0.32	0.32	0.32

Note: This table presents results of the model selection exercise as described in Table 3.2, though this table only includes Abnormal.

Model Selection Without Adjusting for Measurement Error

This section presents the results of the model selection exercise not accounting for measurement error in the human signal. In these analyses, the instruments are constructed using the radiologist's report on the case in the treatment arm without AI assistance. Note that some time elapses between the reads, so the radiologist likely observes a different draw of s_E introducing measurement error into the right-hand side variables of equation 3.7. This is why the preferred method uses instruments constructed using a leave-one-out average of reports for the case. Table C.29 presents results for top-level pathologies with AI, Table C.30 presents results for all pathologies with AI, and Table C.31 presents results for abnormal.

Table C.29: Model selection: top level with AI without accounting for measurement error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.50	0.36	0.58	0.51	0.38	0.36	0.42	0.38	0.48	0.50	0.37	0.37
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
Own information bias (d)	1.00	0.90	0.95	0.94	0.90	0.90	0.93	0.90	0.92	0.94	0.90	0.90
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.34	0.13	0.30	0.28	0.14	0.12	0.21	0.14	0.20	0.28	0.13	0.13
	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	0.61	8.05	0.0	0.0	0.0	8.55	0.04	4.09	0.0	0.08	6.33	5.89
MMSC-BIC	-3.61	-8.82	0.0	0.0	0.0	-8.33	-4.18	-12.79	0.0	-4.14	-10.55	-6.77
Selected moments	4	7	3	3	3	7	4	7	3	4	7	6
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420
R-Squared	0.58	0.55	0.56	0.56	0.55	0.55	0.56	0.55	0.56	0.57	0.55	0.55

Note: This table presents results of the model selection exercise as described in Table 3.2 for top-level pathologies with AI assistance without accounting for measurement error in the radiologist reports.

Table C.30: Model selection: pooled with AI without accounting for measurement error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.46	0.35	0.46	0.44	0.36	0.35	0.39	0.36	0.41	0.43	0.36	0.36
	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Own information bias (d)	1.01	0.94	0.97	0.98	0.94	0.94	0.96	0.94	0.95	0.97	0.95	0.95
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.21	0.00	0.10	0.13	0.02	0.02	0.06	0.01	0.04	0.11	0.03	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	12.06	0.0	0.0	3.42	8.92	10.23	0.0	8.95	0.0	0.0	4.39	7.06
MMSC-BIC	-13.26	0.0	0.0	-0.8	-7.96	-10.86	0.0	-7.93	0.0	0.0	-16.71	-14.04
Selected moments	9	3	3	4	7	8	3	7	3	3	8	8
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100
R-Squared	0.58	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56

Note: This table presents results of the model selection exercise as described in Table 3.2 for all pathologies with AI assistance without accounting for measurement error in the radiologist reports.

Table C.31: Model selection: abnormal without accounting for measurement error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.48	0.36	0.43	0.43	0.36	0.36	0.35	0.37	0.36	0.39	0.36	0.37
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Own information bias (d)	0.94	0.86	0.84	0.88	0.87	0.86	0.84	0.88	0.81	0.85	0.88	0.87
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	1.23	1.01	1.12	1.19	1.01	1.01	0.98	1.04	0.99	1.11	1.03	1.02
	(0.07)	(0.06)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	4.85	8.9	0.0	3.33	7.86	12.55	0.0	10.65	0.0	2.17	8.2	10.85
MMSC-BIC	-7.81	-7.97	0.0	-9.33	-13.24	-12.77	0.0	-14.66	0.0	-2.05	-12.9	-14.46
Selected moments	6	7	3	6	8	9	3	9	3	4	8	9
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710
R-Squared	0.56	0.54	0.53	0.54	0.54	0.54	0.54	0.54	0.52	0.53	0.54	0.54

Note: This table presents results of the model selection exercise as described in Table 3.2 for abnormal without accounting for measurement error in the radiologist reports.

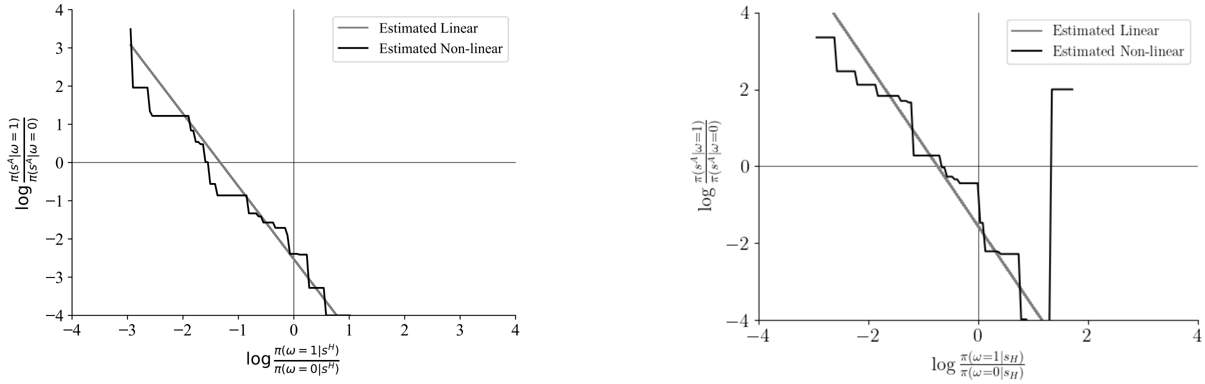
Linearity of the Update Model

To assess the appropriateness of linear relationship in the model of radiologist updating with AI we estimate non-parametric versions of the model and plot the empirical analog of Figure 3.7 along with the joint distribution of signals. To do so, we estimate a boosted tree that estimates the radiologist's reported $\frac{p_h\left(\omega_i=1|s_{ih}^A,s_{ih}^H\right)}{p_h\left(\omega_i=0|s_{ih}^A,s_{ih}^H\right)}$ as a non-parametric function of a constant, the update term, and the reported probability without AI assistance. We impose monotonicity constraints on the update term and the reported probability without AI. We then plot the frontier in which radiologists are indifferent between following up on the case-pathology and not following up. We compare this frontier to the cutoff frontier of a Bayesian decision maker, the radiologist without AI assistance, and the radiologist with AI assistance under the linear model.

Figure C.47: Empirical analog of figure 3.7

(a) Airspace opacity

(b) Cardiomediastinal Abnormality



Note: For the two top-level pathologies with AI assistance, we plot estimates of the indifference frontier where radiologists are indifferent between following up on a patient-case and not following up on a patient-case for a Bayesian decision maker, the linear model estimated in equation 3.7, a non-parametric version of equation 3.7, and the human only without AI assistance. In addition, we plot the joint distribution of the signals log-likelihoods.

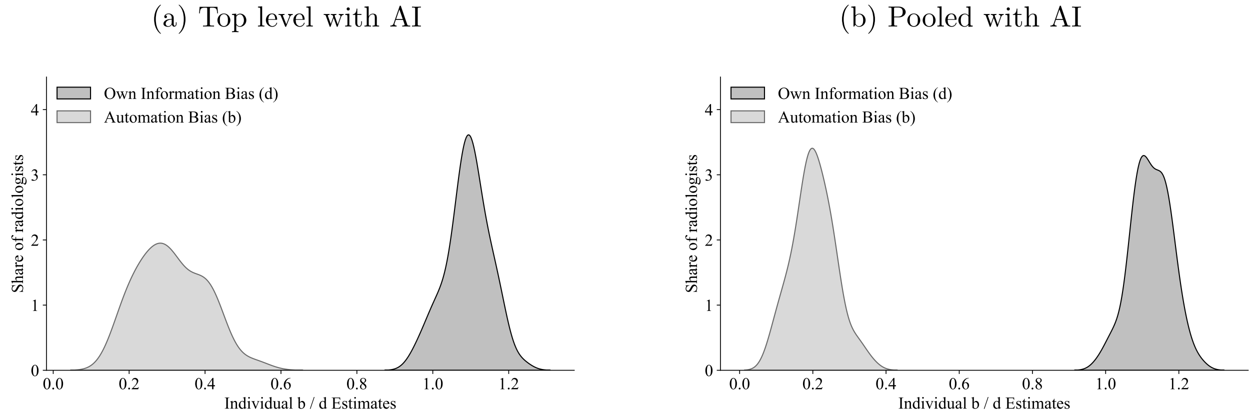
Individual Heterogeneity

Here we show the distribution of individual estimates of equation (3.7). Each estimate contains sampling error, as each radiologist is only reading a subset of cases. Therefore, the unadjusted distribution of individual-level estimates is overdispersed as it is a convolution of the true individual-level parameters and the sampling noise. We adjust for this overdispersion using a Bayesian hierarchical model, where we model the individual parameter vector θ_h as follows.

$$\begin{aligned}
\theta_h &\sim N(\mu, \Sigma) \\
\mu &\sim N(0, 100I) \\
\Sigma &= \text{diag}(\tau)\,\Omega\,\text{diag}(\tau) \\
\tau_k &\sim \text{Cauchy}(0, 2.5) \\
\Omega &\sim \text{LKJCorr}(10)
\end{aligned}$$

We sample from the posterior of this model and plot the marginal distribution of the posterior means of θ_h below.

Figure C.48: Individual heterogeneity in b and d



(a) Top level with AI

(b) Pooled with AI

Note: Marginal distributions of individual b and d estimates by radiologist for top level pathologies with AI and all pathologies with AI.

C.3.7 Preference Estimation

In the experiment we elicit both probability assessment and treatment decisions, allowing us to identify the relative costs of false positives and false negatives the radiologists are using. Recall that radiologist h chooses to treat or follow-up on pathology p in patient case i under treatment t if $a_{hitp} = 1$ where a_{hitp} is given by

$$a_{hitp} = 1\left[\frac{p_{hitp}}{1 - p_{hitp}} - c_{rel}^{hp} + \varepsilon_{hitp} > 0\right]$$

where p_{hitp} is the radiologist's probability assessment, c_{rel}^{hp} is the relative cost of false positives and false negatives for radiologist h and pathology p, and ε_{hitp} captures unobserved preference heterogeneity. If ε_{hitp} follows a Logistic distribution, we can estimate c_{rel}^{hp} through a logistic regression. We impose a low-dimensional structure on c_{rel}^{hp} to improve statistical precision and estimate the following logistic regression

$$\log \frac{P(a_{hitp}=1)}{1-P(a_{hitp}=1)} = \beta_0 + \beta \log \frac{p_{hitp}}{1-p_{hitp}} + \alpha_p + \gamma_h \tag{C.1}$$

where α_p are pathology fixed effects and γ_h are radiologist fixed effects. The relative costs of false positives to false negatives for radiologist h and pathology p can then be found as $c_{rel}^{hp} = \exp\left[-\frac{\beta_0+\gamma_h+\alpha_p}{\beta}\right]$. For each pathology, we winsorize radiologists' relative costs at the 5th and 95th percentile. The results of this exercise are presented in Table C.32.

Table C.32: Preference estimates

	Mean	Std	Min	25%	50%	75%	Max
All	3.696	5.278	0.243	0.663	1.513	3.913	20.884
Top (with AI)	1.164	1.522	0.155	0.271	0.489	1.263	5.849
AI	2.139	3.058	0.193	0.383	0.891	2.179	12.287
Abnormal	2.223	2.952	0.331	0.519	0.966	2.379	11.667

Note: Distribution of c_{rel}^{hp} for each of the four pre-registered pathology groups calculated from the estimates of equation C.1. The distribution of c_{rel}^{hp} is winsorized for each pathology at the 5th and 95th percentile.

C.3.8 Delegation Results for Cardiomediastinal Abnormality

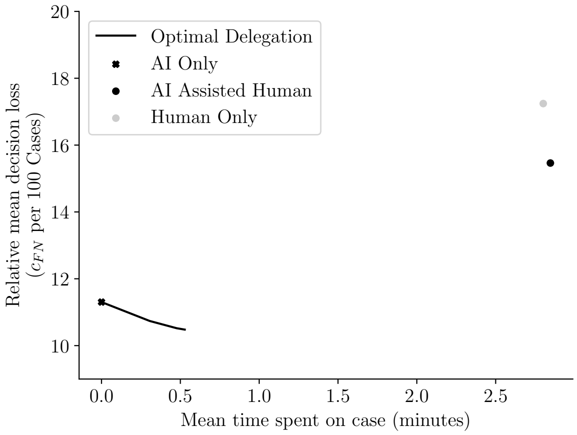
Here we show the results of Section 3.6 for the other top-level pathology with AI assistance, Cardiomediastinal Abnormality. Table C.33 shows the decision loss and time cost for various delegation strategies, Figure C.49 plots the possibilities frontier between human time and decision loss, and Figure C.50 plots the share of cases assigned to each modality under the optimal delegation strategy for a range of values of the social cost of a false negative.

Table C.33: Cardiomediastinal Abnormality delegation results

	Time Cost		Pr(Fp)	Pr(Fn)	Decision Loss
	Minutes	Dollars			
Bayesian	2.8	11.4	3.2	6.3	8.4
AI Only	0.0	0.0	4.4	7.6	10.5
Human Only	2.8	11.2	20.5	3.6	17.2
Human + AI	2.8	11.4	18.4	3.2	15.5
Min. Decision Loss	0.5	2.1	3.3	6.4	10.5

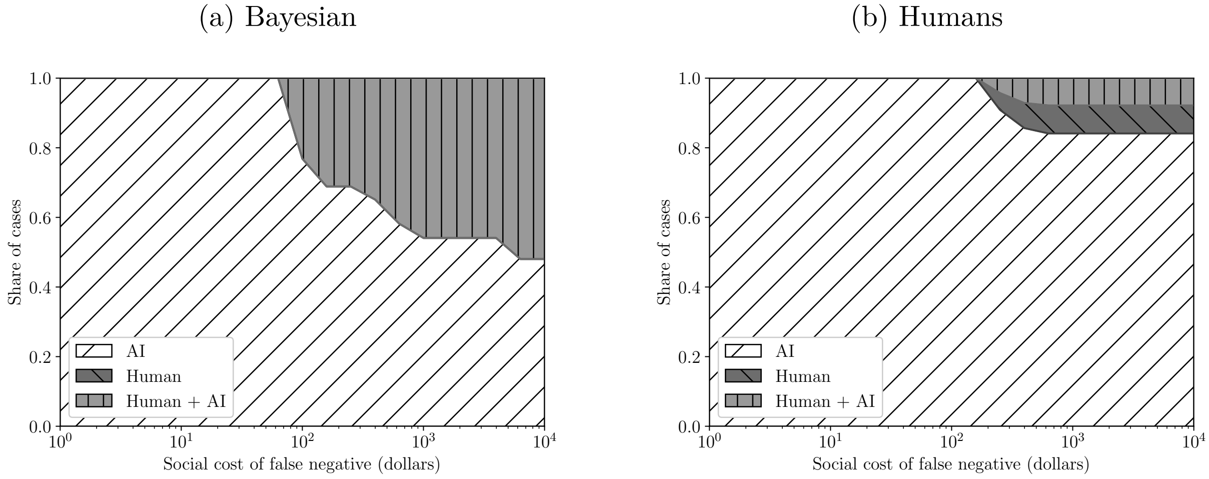
Note: This table shows the time taken and decision loss of various delegation strategies for Cardiomediastinal Abnormality. The average time per case is shown in both minutes and dollars using a wage of $4 per minute. The table also reports the share of false positives ($Pr(FP)$), the share of false negatives ($Pr(FN)$), and decision loss calculated as $Pr(FN) + c_{rel}Pr(FP)$ where c_{rel} =0.66 – the median c_{rel} for Cardiomediastinal Abnormality. The Bayesian row shows results for the Bayesian decision maker. AI Only shows results for full delegation to the AI. Human Only shows results if humans read cases on their own without AI assistance. Human + AI shows results if humans read all cases with access to the AI. Min. Decision Loss shows results for the optimal delegation strategy that minimizes decision loss and highlights the potential improvement in decisions from delegating to the AI. This analysis excludes data from design 3 because of learning effects in this setup.

Figure C.49: Loss-time frontier: Cardiomediastinal Abnormality



Note: This graph shows how human radiologists and the AI perform relative to the optimal delegation system on the frontier of the cost of human time versus decision loss. This analysis excludes data from design 3 because of learning effects in this setup.

Figure C.50: Cardiomediastinal Abnormality modality shares


(a) Bayesian

(b) Humans

Note: The graphs show the share of cases decided by each modality (humans, AI, humans+AI) conditional on the cost of a false negative in dollars, denoted m in the text, for airspace opacity. Panel (a) focuses on a Bayesian decision maker. Panel (b) focuses on a human decision-maker with decisions and time-taken as in our experiment. This analysis excludes data from design 3 because of learning effects in this setup.

References

Alberto Abadie and Matias D Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503, 2018.

Jason Abaluck, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. The determinants of productivity in medical testing: Intensity and allocation of care. *Am. Econ. Rev.*, 106 (12):3730–3764, December 2016.

Daron Acemoglu and Simon Johnson. Power and progress: Our Thousand-Year struggle over technology and prosperity. *Public Affairs, New York*, 2023.

Amanda Y Agan, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Working Paper 30981, National Bureau of Economic Research, February 2023. URL http://www.nber.org/papers/w30981.

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, 2023.

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction Machines: The Simple Economics of Artificial Intelligence.* Harvard Business Press, April 2018.

Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. Artificial intelligence: The ambiguous labor market impact of automating prediction. *J. Econ. Perspect.*, 33(2):31–50, May 2019.

Wajeeha Ahmad, Ananya Sen, Charles Eesley, and Erik Brynjolfsson. The role of advertisers and platforms in monetizing misinformation: Descriptive and experimental evidence. Technical report, Working Paper, 2023.

Jong Seok Ahn, Shadi Ebrahimian, Shaunagh McDermott, Sanghyup Lee, Laura Naccarato, John F Di Capua, Markus Y Wu, Eric W Zhang, Victorine Muse, Benjamin Miller, Farid Sabzalipour, Bernardo C Bizzo, Keith J Dreyer, Parisa Kaviani, Subba R Digumarthy, and Mannudeep K Kalra. Association of artificial Intelligence–Aided chest radiograph interpretation with reader performance and efficiency. *JAMA Netw Open*, 5(8):e2229289–e2229289, August 2022.

Eugenio Alberdi, Lorenzo Strigini, Andrey A Povyakalo, and Peter Ayton. Why are people's decisions sometimes worse with computer support? In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 18–31. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. The Welfare Effects of Social Media. *American Economic Review*, 110, 2020. ISSN 0002-8282. doi:10.1257/aer.20190658.

Hunt Allcott, Matthew Gentzkow, and Lena Song. Digital addiction. *American Economic Review*, 112(7):2424–63, 2022.

Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*, pages 2155–2165, 2020.

Donald W K Andrews and Biao Lu. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J. Econom.*, 101(1): 123–164, March 2001.

Victoria Angelova, Will Dobbie, and Crystal Yang. Algorithmic recommendations and human discretion. 2022.

Guy Aridor, Duarte Gonçalves, Daniel Kluver, Ruoyan Kong, and Joseph Konstan. The economics of recommender systems: Evidence from a field experiment on movielens. *arXiv preprint arXiv:2211.14219*, 2022.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015a.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, June 2015b. ISSN 0036-8075, 1095-9203. doi:10.1126/science.aaa1160.

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate AI the best teammate? optimizing AI for teamwork. *AAAI*, 35(13):11405–11414, May 2021.

Nicholas Barberis, Andrei Shleifer, and Robert Vishny. A model of investor sentiment. *J. financ. econ.*, 49(3):307–343, September 1998.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn.

The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

Dan Benjamin, Aaron Bodoh-Creed, and Matthew Rabin. Base-rate neglect: Foundations and implications. 2019.

Daniel J Benjamin. Chapter 2 - errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 69–186. January 2019.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *J. Econ. Lit.*, 57(1):44–95, March 2019.

Ron Berman and Zsolt Katona. Curation algorithms and filter bubbles in social networks. *Marketing Science*, 39(2):296–316, 2020.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

Steven T Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262, 1994.

David Blackwell. Equivalent comparisons of experiments. *Ann. Math. Stat.*, 24(2):265–272, 1953.

Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. 358(6370):1530–, 2017.

Erik Brynjolfsson, Daniel Rock, and Chad Syverson. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. November 2017.

Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.

Kate Bundorf, Maria Polyakova, and Ming Tai-Seale. How do humans interact with algorithms? experimental evidence from health insurance. June 2020.

Moira Burke, Cameron Marlow, and Thomas Lento. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 945–954, 2009.

Gordon Burtch, Yili Hong, Ravi Bapna, and Vladas Griskevicius. Stimulating online reviews by combining financial incentives and social norms. *Management Science*, 64(5):2065–2082, 2018.

Gordon Burtch, Qinglai He, Yili Hong, and Dokyun Lee. How do peer awards motivate creative content? experimental evidence from reddit. *Management Science*, 2021.

Luis Cabral and Lingfang Li. A dollar for your thoughts: Feedback-conditional rebates on ebay. *Management Science*, 61(9):2052–2063, 2015.

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, 2020. doi:10.1017/9781108684606.

Matias D. Cattaneo, Nicolas Idrobo, and Rocio Titiunik. A practical introduction to regression discontinuity designs: Extensions, 2023.

David C Chan, Matthew Gentzkow, and Chuan Yu. Selection with variation in diagnostic skill: Evidence from radiologists. *Q. J. Econ.*, 137(2):729–783, May 2022.

Amitabh Chandra and Douglas O Staiger. Identifying sources of inefficiency in healthcare. *Q. J. Econ.*, 135(2):785–843, May 2020.

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.

Daniel L Chen, Martin Schonger, and Chris Wickens. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, March 2016.

Guangying Chen, Tat Chan, Dennis Zhang, Senmao Liu, and Yuxiang Wu. The effects of diversity in algorithmic recommendations on digital content consumption: A field experiment. *Available at SSRN 4365121*, 2023.

Yan Chen, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, 100(4):1358–98, 2010.

Jörg Claussen, Christian Peukert, and Ananya Sen. The editor vs. the algorithm: Returns to data and externalities in online news. *Working Paper*, 2019.

Jörg Claussen, Christian Peukert, and Ananya Sen. The editor and the algorithm: Returns to data and externalities in online news. *Available at SSRN 3479854*, 2021.

Emily F Conant, Alicia Y Toledano, Senthil Periaswamy, Sergei V Fotin, Jonathan Go, Justin E Boatsman, and Jeffrey W Hoffmeister. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell*, 1(4):e180096, July 2019.

John J Conlon, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach. Not learning from others. August 2022.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

Janet Currie and W Bentley MacLeod. Diagnosing expertise: Human capital, decision making, and performance among physicians. *J. Labor Econ.*, 35(1), 2017.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.*, 144(1):114–126, February 2015.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manage. Sci.*, 64(3):1155–1170, March 2018.

Yingying Dong and Michal Kolesár. When can we ignore measurement error in the running variable? *arXiv preprint arXiv:2111.07388*, 2021.

Robert Donnelly, Ayush Kanodia, and Ilya Morozov. Welfare effects of personalized rankings. *Marketing Science*, 2023.

Jack Dorsey. They simply try to put the tweets that you're *most likely* to engage with at the top... https://twitter.com/jack/status/1525662031440924672, May 2022. Tweet.

Erwan Dujeancourt, Marcel Garz, Anindya Ghose, Johannes Hagen, Juliane Lischka, Mattias Nordin, Jonna Rickardsson, and Marco Schwarz. The effects of algorithmic content selection on user engagement with news on twitter. Technical report, Working Paper, 2021.

Dean Eckles. Algorithmic transparency and assessing effects of algorithmic ranking. 2022.

Dean Eckles, René F Kizilcec, and Eytan Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27): 7316–7322, 2016.

Benjamin Enke and Florian Zimmermann. Correlation neglect in belief formation. *The Review of Economic Studies*, 86(1):313–332, 2019.

Ed Felten, Manav Raj, and Robert Seamans. How will language modelers like ChatGPT affect occupations and industries? March 2023.

Edward W Felten, Manav Raj, and Robert Seamans. The occupational impact of artificial intelligence: Labor, skills, and polarization. September 2019.

Daniel Fleder and Kartik Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, 55(5):697–712, 2009.

Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? influences of Human-AI workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1362–1374. Association for Computing Machinery, June 2022.

Morgan R Frank, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan. Toward understanding the impact of artificial intelligence on labor. *Proc. Natl. Acad. Sci. U. S. A.*, 116(14):6531–6539, April 2019.

Devin Gaffney and J Nathan Matias. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PloS one*, 13(7):e0200162, 2018.

Jana Gallus. Fostering public good contributions with symbolic awards: A large-scale natural field experiment at wikipedia. *Management Science*, 63(12):3999–4015, 2017.

Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo Koch, Matthias Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, Felipe Kitamura, Marzyeh Ghassemi, and Errol Colak. Who should do as AI say? only non-task expert physicians benefit from correct explainable AI advice. June 2022.

Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias F C Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, Felipe C Kitamura, Marzyeh Ghassemi, and Errol Colak. Non-task expert physicians benefit from correct explainable AI advice when reviewing x-rays. *Sci. Rep.*, 13(1):1383, January 2023.

Paul J Gertler, Sebastian Martinez, Patrick Premand, Laura B Rawlings, and Christel MJ Vermeersch. *Impact evaluation in practice.* World Bank Publications, 2016.

Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science*, 60(7):1632–1654, 2014.

Paulo B Goes, Chenhui Guo, and Mingfeng Lin. Do incentive hierarchies induce user effort? evidence from an online knowledge exchange. *Information Systems Research*, 27(3):497–516, 2016.

Avi Goldfarb, Bledi Taska, and Florenta Teodoridis. Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. *Res. Policy*, 52(1):104653, January 2023.

David M Grether. Bayes rule as a descriptive model: The representativeness heuristic. *Q. J. Econ.*, 95(3):537–557, November 1980.

David M Grether. Testing bayes rule and the representativeness heuristic: Some experimental evidence. *J. Econ. Behav. Organ.*, 17(1):31–57, January 1992.

Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cogn. Psychol.*, 24(3):411–435, July 1992.

Grimon, Marie-Pascale, and Christopher Mills. The impact of algorithmic tools on child protection: Evidence from a randomized controlled trial. 2022.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

Jonathan Gruber, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad. Managing intelligence: Skilled experts and decision support in markets for complex products. 2021.

Bin Gu, Prabhudev Konana, Balaji Rajagopalan, and Hsuan-Wei Michelle Chen. Competition among virtual communities and user valuation: The case of investing-related communities. *Information systems research*, 18(1):68–85, 2007.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1):eaau4586, 2019.

Andrew M Guess, Brendan Nyhan, and Jason Reifler. Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, 4(5):472–480, 2020.

Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023.

H Benjamin Harvey and Vrushab Gowda. How the FDA regulates AI. *Acad. Radiol.*, 27(1): 58–61, January 2020.

Frances Haugen. Statement of frances haugen. *United States Senate Committee on Commerce, Science and Transportation*, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

David Holtz, Ben Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral. The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 75–76, 2020.

Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823, 2014.

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo J W L Aerts. Artificial intelligence in radiology. *Nat. Rev. Cancer*, 18(8):500–510, August 2018.

Tanjim Hossain and Ryo Okui. The binarized scoring rule. *Rev. Econ. Stud.*, 80(3):984–1001, February 2013.

Tiffany Hsu and Gillian Friedman. Facebook boycott: Starbucks and diageo to pull ads. *The New York Times*, 2020. URL https://www.nytimes.com/2020/06/26/business/media/Facebook-advertising-boycott.html.

Tiffany Hsu and Eleanor Lutz. More than 1,000 companies boycotted facebook. did it work? *The New York Times*, 2020. URL https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. Ieee, 2008.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *arXiv preprint arXiv:2110.11010*, 2021.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022.

Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24, 2012.

Kosuke Imai, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of Algorithm-Assisted human Decision-Making: Application to pretrial public safety assessment. December 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, July 2019.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035, May 2016.

Daniel Kahneman. *Thinking, fast and slow.* macmillan, 2011.

Daniel Kahneman and Amos Tversky. On the psychology of prediction. *Psychol. Rev.*, 80 (4):237–251, July 1973.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *Am. Econ. Rev.*, 101(6): 2590–2615, October 2011.

Simon Kemp. Digital 2020 reports. *"wearesocial.com/digital-2020*, 2020.

Warut Khern-am nuai, Karthik Kannan, and Hossein Ghasemkhani. Extrinsic versus intrinsic rewards for contributing reviews in an online platform. *Information Systems Research*, 29(4):871–892, 2018.

Hyo Eun Kim, Hak Hee Kim, Boo Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun Kyung Kim. Changes in cancer detection and False-Positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2(3):e138–e148, March 2020.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *Am. Econ. Rev.*, 105(5):491–495, May 2015.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *Q. J. Econ.*, 133(1):237–293, August 2017.

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

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

Barnett S Kramer, Christine D Berg, Denise R Aberle, and Philip C Prorok. Lung cancer screening with low-dose helical CT: results from the national lung screening trial (NLST). *J. Med. Screen.*, 18(3):109–111, 2011.

Akos Lada, Meihong Wang, and Tak Yan. How machine learning powers facebook's news feed ranking algorithm. *Facebook Engineering*, 2021.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of Human-AI decision making: A survey of empirical studies. December 2021.

Curtis P Langlotz. Will artificial intelligence replace radiologists? *Radiology: Artificial Intelligence*, 1(3):e190058, May 2019.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

Dokyun Lee and Kartik Hosanagar. How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research*, 30(1):239–259, 2019.

Ro'ee Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70, 2021.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David Rand, and Gordon Pennycook. High level of agreement across different news domain quality ratings. 2022.

John DC Little. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, October 2019.

Robert Mccluskey, A Enshaei, and B A S Hasan. Finding the Ground-Truth from multiple labellers: Why parameters of the task matter. *ArXiv*, 2021.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR, 2020.

Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

S Mullainathan and Z Obermeyer. A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions. 2019.

Sendhil Mullainathan and Ziad Obermeyer. Does machine learning automate moral hazard and error? *American Economic Review*, 107(5):476–480, 2017.

Simha Mummalaneni, Hema Yoganarasimhan, and Varad V Pathak. Producer and consumer engagement on social media platforms. *Available at SSRN 4173537*, 2022.

Arvind Narayanan. Understanding social media recommendation algorithms. *Knight First Amendment Institute*, 2023. URL https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.

Sridhar Narayanan and Kirthi Kalyanam. Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407, 2015.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, 2020.

Justin G Norden and Nirav R Shah. What AI in health care can learn from the long road to autonomous vehicles. *NEJM Catalyst Innovations in Care Delivery*, 3(2), 2022.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. March 2023.

Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future — big data, machine learning, and clinical medicine. *N. Engl. J. Med.*, 375(13):1216–1219, September 2016.

Gal Oestreicher-Singer and Arun Sundararajan. Recommendation networks and the long tail of electronic commerce. *MIS Quarterly*, pages 65–83, 2012.

Gal Oestreicher-Singer and Lior Zalmanson. Content or community? a digital business strategy for content providers in the social age. *MIS quarterly*, pages 591–616, 2013.

Jeff Orlowski. The social dilemma. Netflix, 2020. URL https://www.thesocialdilemma.com/.

Serena Pacilè, January Lopez, Pauline Chone, Thomas Bertinotti, Jean Marie Grouin, and Pierre Fillard. Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiol Artif Intell*, 2(6):e190208, November 2020.

David M Panicek and Hedvig Hricak. How sure are you, doctor? a standardized lexicon to describe the radiologist's level of certainty. *AJR Am. J. Roentgenol.*, 207(1):2–3, July 2016.

Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin, 2011.

Allison Park, Chris Chute, Pranav Rajpurkar, Joe Lou, Robyn L Ball, Katie Shpanskaya, Rashad Jabarkheel, Lily H Kim, Emily McKenna, Joe Tseng, Jason Ni, Fidaa Wishah, Fred Wittber, David S Hong, Thomas J Wilson, Safwan Halabi, Sanjay Basu, Bhavik N Patel, Matthew P Lungren, Andrew Y Ng, and Kristen W Yeom. Deep Learning-Assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA network open*, 2(6): e195600, June 2019.

Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A J Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew Lungren. Human–Machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine*, 2(1):111, December 2019.

Rajan Patel. Google expands partnership with reddit, 2024. URL https://blog.google/inside-google/company-announcements/expanded-reddit-partnership/.

Gordon Pennycook and David G Rand. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402, 2021.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462*, 2021.

M Rabin. Inference by believers in the law of small numbers. *Q. J. Econ.*, 2002.

M Rabin and D Vayanos. The gambler's and hot-hand fallacies: Theory and applications. *Rev. Econ. Stud.*, 2010.

Matthew Rabin. Incorporating limited rationality into economics. *J. Econ. Lit.*, 51(2): 528–543, June 2013.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv*, March 2019.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng. CheXNet: Radiologist-Level pneumonia detection on chest X-Rays with deep learning. (1711.05225), December 2017.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Francis G Blankenberg, Jayne Seekins, Timothy J Amrhein, David A Mong, Safwan S Halabi, Evan J Zucker, Andrew Y Ng, and Matthew P Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.*, 15(11):e1002686, November 2018.

Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, Rulan Griesel, Andrew Y Ng, Tom H Boyles, and Matthew P Lungren. CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest X-Rays in patients with HIV. *npj Digital Medicine*, 3:115, December 2020.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. AI in health and medicine. *Nat. Med.*, 28(1):31–38, January 2022.

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

reddit.com. Reddit, 11 2017. URL https://github.com/reddit-archive/reddit.

Michael Restivo and Arnout Van De Rijt. Experimental study of informal rewards in peer production. *PloS one*, 7(3):e34358, 2012.

Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.*, 12(1):1–10, September 2022.

Stephen A Rhoades. The herfindahl-hirschman index. *Fed. Res. Bull.*, 79:188, 1993.

Michael Allan Ribers and Hannes Ullrich. Machine predictions and human decisions with variation in payoff and skills: the case of antibiotic prescribing. 2022.

Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

Andrew B Rosenkrantz, Tarek N Hanna, Scott D Steenburg, Mary Jo Tarrant, Robert S Pyatt, and Eric B Friedberg. The current state of teleradiology across the united states: A national survey of radiologists' habits, attitudes, and perceptions on teleradiology practice. *J. Am. Coll. Radiol.*, 16(12):1677–1687, December 2019.

Jarrel C Y Seah, Cyril H M Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, Ben Hachey, Stephen J F Hogg, Benjamin P Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine M Jones. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health*, 3(8):e496–e506, August 2021.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 614–622, New York, NY, USA, August 2008. Association for Computing Machinery.

Yongsik Sim, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, Synho Do, Kyunghwa Han, Hanmyoung Kim, Seungwook Yang, Dong-Jae Lee, and Byoung Wook Choi. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology*, 294(1):199–209, January 2020.

SimilarWeb. Top social media networks websites ranking in march 2024, 2024. URL https://www.similarweb.com/top-websites/computers-electronics-and-technology/social-networks-and-online-communities/.

Christopher A Sims. Implications of rational inattention. *J. Monet. Econ.*, 50(3):665–690, April 2003.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. April 2020.

Michael Spence and Bruce Owen. Television programming, monopolistic competition, and welfare. *The Quarterly Journal of Economics*, 91(1):103–126, 1977.

Megan Stevenson and Jennifer L Doleac. Algorithmic risk assessment in the hands of humans. December 2019.

Cass Sunstein. R. 2007. republic. com 2.0, 2003.

Yasasvi Tadavarthi, Brianna Vey, Elizabeth Krupinski, Adam Prater, Judy Gichoya, Nabile Safdar, and Hari Trivedi. The state of radiology AI: Considerations for purchase decisions and current market offerings. *Radiol Artif Intell*, 2(6):e200004, November 2020.

Matt Taddy. The technological elements of artificial intelligence. February 2018.

Luke Thorburn, Priyanjana Bengani, and Jonathan Stray. How platform recommenders work. *Medium*, 2022. URL https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nat. Med.*, 26(8):1229–1234, June 2020.

A Tversky and D Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, September 1974.

Twitter. Twitter's recommendation algorithm. *Twitter Engineering Blog*, 2023. URL https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

Raluca M Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018.

Marshall Van Alstyne and Erik Brynjolfsson. Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6):851–868, 2005.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018a.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The Spread of True and False News Online. *Science*, 359(6380):1146–1151, 2018b.

Thomas S Wallsten and Adele Diederich. Understanding pooled subjective probability estimates. *Math. Soc. Sci.*, 41(1):1–18, January 2001.

Yuyan Wang, Long Tao, and Xing Zhang. Recommending for a three-sided food delivery marketplace: A multi-objective hierarchical approach. Technical report, Working Paper, 2023.

Mark R. Warner. Warner presses meta on facebook's role in inciting violence and spreading misinformation around the world. Press Release, 2023.

Michael Webb. The impact of artificial intelligence on the labor market. 158713(November), 2019.

Stefan Wojcik and Adam Hughes. Sizing up twitter users, April 2019. URL https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE*, 109(5):820–838, May 2021.