Machine Predictions and Causal Explanations: Evidence from a Field Experiment

Xi Kang and Hyunjin Kim *

June 30, 2025

Abstract

A key role that decision-makers play in organizations is to provide causal explanations for decisions. Yet despite growing evidence on how machine predictions impact decisions, there has been less insight on how they affect decision-makers' ability to explain and reason through their decisions. We explore this question in this paper, using a field experiment we designed and ran across mutual fund analysts in a leading investment research firm. We randomly assigned analysts to receive predictions for fund ratings based on a proprietary machine learning algorithm developed within the company, and observed their decisions and causal explanations for those decisions over 6 months. We find that even when predictions improve decisions, they can worsen the reasoning of the causal explanations that decision-makers provide for their decisions. Inexperienced analysts working on simpler decisions are especially affected, leading them to provide worse explanations when working with machine predictions. These findings suggest that the growing use of AI may accrue knowledge without understanding, especially hindering inexperienced decision-makers from developing expertise.

*Authors contributed equally and are listed alphabetically. Xi Kang: Vanderbilt University, xi.kang@vanderbilt.edu; Hyunjin Kim: INSEAD, hyunjin.kim@insead.edu. This experiment would not have been possible without the help of our collaborating company and key support from Cynthia Pekron. We thank Ruijing Chen, Can He, and Xinni Wu for their excellent research assistance. We are grateful for valuable comments from Arnaldo Camuffo, Giada Di Stefano, Dan Gross, Phanish Puranam, Victoria Sevcenko, Jasjit Singh, and seminar participants at MIT, HBS, University of Maryland, University of Minnesota, Ohio State University, UC Irvine, the Strategy Research Forum, Bocconi ION lab, HEC Paris AI and Entrepreneurship Workshop, the MAD conference, Wharton AI and the Future of Work conference, Strategy Science, DRUID, and Academy of Management. We are grateful for financial support from INSEAD and Vanderbilt University. This experiment was approved by the INSEAD IRB office and pre-registered at the AEA RCT Registry. All errors are our own.

1 Introduction

Firms are increasingly using machine predictions and AI systems to guide strategic decisions across a wide range of domains, including hiring (Cowgill 2019, Dell'Acqua 2022), resource allocation (Kim et al., 2024), and innovation (Agrawal et al. 2018, Lou and Wu 2021). Growing empirical evidence suggests that machine predictions can improve decisions through superior forecasting accuracy, with some studies highlighting the importance of reducing "algorithm aversion" to increase decision-makers' usage of algorithmic predictions (Dietvorst et al. 2015, Allen and Choudhury 2022, Lebovitz et al. 2022, Tong et al. 2021).

While considerable attention has been paid to how machine predictions influence decision quality, less is known about how they affect decision-makers' ability to construct persuasive, well-reasoned causal explanations—a core managerial task (Felin and Zenger 2009, 2017, Helfat and Peteraf 2015, Mercier and Sperber 2011, Sørensen and Carroll 2021). Entrepreneurs, for instance, must articulate compelling narratives to secure funding (Martens et al. 2007). Managers must justify strategic choices to motivate employees and persuade stakeholders.[1] These explanations rely on mental models shaped by experience and expertise, which support decision-making under uncertainty and stakeholder persuasion (Barron and Fries 2024, Angrisani et al. 2024, Kendall and Charles 2022, Eliaz and Spiegler 2020, Schoar and Sun 2024) . While machine predictions may enhance short-run accuracy (Agrawal et al. 2018, Camuffo et al. 2022), they may degrade the quality of causal explanations behind decisions and hinder the development of expertise. This has long-term implications for organizations, particularly as human judgment becomes critical in edge cases where algorithms fail due to bias, overfitting, or noise (Shrestha et al. 2019, Choudhury et al. 2020, Raisch and Krakowski 2021).

In this paper, we propose that machine predictions affect how decision-makers explain the decisions they make, separate from their impact on the decisions themselves. We find that even when algorithms improve decisions, they can make the causal reasoning that decision-makers provide for the decision less clear and persuasive (as well as the reverse). Less experienced decision-makers are more impacted, showing worse reasoning on simpler decisions when relying on machine predictions – even though these decisions improve. We find suggestive evidence that these results are driven by less experienced analysts being less able to reason and construct a narrative when relying on machine predictions. These findings have important implications for organizational learning and

[1] A prominent example is Amazon's six-page memo, described as "the way Amazon makes [major] decisions," requiring a narrative rationale (Stewart 2020).

expertise development, highlighting a potential cost of delegating predictive tasks to machines.

We designed and ran a field experiment for six months across all fund analysts in a leading financial research firm, which provided a unique opportunity to examine the impact of predictive algorithms across decision-makers of varying experience on different types of decisions. Over a 9-month period, we observed all decisions made by analysts, who conducted research on mutual funds worldwide and provided forward-looking ratings and reports for institutional and retail investors. Analysts performed their work on an online interface where they could see their assigned fund's rating from the last evaluation and receive information about the fund for their analysis. They decided on an up-to-date rating for the fund and provided a causal explanation for their rating decision, which was assessed by an expert review committee within the company before being finalized and released.

At the time of the experiment, the company had developed a proprietary machine learning algorithm that provided predictions on mutual fund ratings, which was representative of the most prevalent form of AI adoption in the finance industry.[2] By providing machine predictions, the company sought to shift analyst effort into improving their causal explanations, which were generally valued more highly than the ratings decisions – by investors who were their clients, as well as other financial research companies who employed analysts. These explanations were sought out from investors to help them develop a forward-looking outlook about the fund based on the analysts' unique perspectives and contextual insights.

We randomly assigned analysts to control or treatment conditions. Treated analysts received a machine-generated fund rating alongside the standard information interface. Half of the treated analysts also received an explainability feature: a Shapley value-based visualization highlighting the fund characteristics that contributed most to the machine's prediction – which was at the frontier of the computer science literature at the time (Lundberg and Lee 2017). We first examine how machine predictions affected analysts' likelihood of changing the rating, as a measure of whether algorithms had any effect on their decision-making. We evaluate whether they improved analysts' decisions by assessing whether treated analysts' recommended funds performed better than those

[2]The company developed a proprietary machine learning algorithm using ensemble random forest models, with continuous backtesting since 2015 for performance validation. Prior to our experiment, the algorithm development team documented methodology and performance in internal reports and conducted training sessions to ensure analysts understood the approach. Importantly, analysts were fully aware of the algorithm's quality before the experiment began. The internal performance documentation showed that across all fund types, highly-recommended funds outperformed their category average by approximately 1%, while not-recommended funds underperformed by 1% over 12-month periods. This consistent performance across fund types was well-known to analysts through the training process.

in the control group in the subsequent 3-6 months. We then examine the impact of algorithms on analysts' causal explanations using two approaches: (1) expert committee ratings of report quality and (2) text-based analysis of explanations using natural language processing. We also validate these internal expert committee ratings through a follow-up study with external financial advisors who are the firm's customers of the analyst reports and collectively manage $4.8 billion in total assets. We find that their assessments significantly correlate with internal committee evaluations across all dimensions, and that they indicate greater willingness to invest in funds that receive higher committee ratings, conditional on the same rating.

We find that machine predictions significantly increased the likelihood that analysts changed both their rating decisions and the causal explanations in their reports. However, we observe no precise treatment effects on average on their decision or explanation quality. We find substantial heterogeneity in treatment effects, suggesting that machine predictions impacted decisions and reasoning in opposite directions across decision types, especially for less experienced analysts. For new analysts, machine predictions improved the accuracy of simpler (NAE) fund ratings but worsened the quality of their accompanying explanations. For more complex (AE) funds, the opposite pattern emerged: new analysts made less accurate decisions but provided more coherent causal explanations. Experienced analysts were less affected by the intervention overall.

To understand the mechanisms behind these effects, we analyze the texts of analysts' causal explanations, their perceptions of machine predictions, and their responses to algorithm explainability. We present suggestive evidence that analysts face a cognitive constraint in incorporating machine predictions: they must trade off between accepting a prediction they cannot fully rationalize or defaulting to their own predictions and explanatory models. Because machine predictions outperform analysts' own forecasts on average, this results in a tradeoff between decision and reasoning quality. Relying on machine predictions improves predictive accuracy but degrades their causal explanations, while relying on one's own model preserves narrative coherence at the cost of decision quality.

This tradeoff is most acute for less experienced analysts. We find that they are more likely to rely on machine predictions for simpler decisions—where prediction quality improves but explanation quality deteriorates. In these cases, their reports are shorter, less technical, and less coherent, suggesting limited ability to reason effectively when outsourcing prediction. By contrast, they rely less on machine predictions for complex decisions and instead lean on their own explanatory models. Qualitative evidence suggests that this prompted them to argue against the machine—challenging

specific Shapley-derived drivers and defending alternative interpretations, which may have improved reasoning. Experienced analysts, in contrast, were better able to integrate machine predictions and selectively incorporate machine-suggested features to sharpen their explanations. Taken together, these findings suggest that machine predictions can improve decisions but also introduce a cognitive burden on less experienced decision-makers, potentially undermining their ability to construct coherent causal narratives—an essential aspect of expertise development, stakeholder persuasion, and long-run organizational learning.

Our study contributes to growing research on algorithms and decision-making in organizations, which has provided insight on how the use of algorithms affects firm decisions and performance (Choudhury et al. 2020, Kim et al. 2024, Tong et al. 2021, Brynjolfsson et al. 2021). In this study, we highlight how the use of algorithms may also affect how decision-makers reason about and explain decisions, and suggest that algorithms may worsen their causal explanations even when improving decisions. Unlike human decision-making — where causal reasoning often precedes and evolves alongside the act of making a decision — machine predictions emerge from opaque, correlational processes that analysts struggle to integrate into coherent causal narratives. This creates a fundamental tradeoff: algorithms may improve decisions while simultaneously deteriorating the causal explanations that decision-makers provide, especially for those who are less experienced. These findings suggest that organizations may need to be careful about using algorithms to support decisions, even if they yield gains in the short-term.

Our study contributes to research on the critical role of domain expertise in enabling decision-makers to engage with algorithmic outputs and the implications for skill development (Choudhury et al. 2020, Lebovitz et al. 2022). We find that experienced analysts are better equipped to incorporate machine predictions, contextualize them within broader causal models, and integrate them into coherent explanations. However, our results raise the possibility that widespread reliance on algorithmic predictions may undermine the very processes through which such expertise is developed. The act of crafting causal explanations—forming hypotheses, interpreting outcomes, and revising mental models—is not merely ancillary to decision-making but foundational to learning and long-term skill development. When individuals rely on machine outputs without being able to integrate them into their own causal reasoning, they may accumulate knowledge about what works without understanding why it works, presenting challenges for organizational learning and expertise development even when decisions improve in the short-run.

Finally, our study contributes to emerging research on explainable AI by providing empirical

evidence of a fundamental tension identified in the literature on human-AI collaboration. Algorithms optimize their objectives based on correlations – a gap that becomes particularly problematic in high-stakes settings like investment decisions where experts need to provide causal explanations to stakeholders (Burkart and Huber 2021). We find suggestive evidence that algorithm explainability may indeed help increase the usage of algorithms, even when it does not provide an accurate or consistent explanation behind machine predictions. Our results suggest that explainability may be a double-edged sword, as it can encourage the adoption of predictive algorithms by decision makers who may be averse to machine predictions (Dietvorst et al. 2015, Allen and Choudhury 2022), while simultaneously risking undermining their ability to construct coherent causal narratives, particularly for less experienced users. Our findings could be driven by the particular feature attribution method of explainability used in our setting, which is inherently based on correlation. As the explainable AI literature continues to develop technical solutions for model transparency, future studies should investigate how alternative explainability methods can support decision-makers in overcoming the fundamental cognitive challenge of integrating machine predictions with human causal reasoning.

2 Causal explanations and machine predictions

2.1 The importance of causal explanations in strategic decision-making

Communicating causal explanations for decisions is a key component of strategic decision-making. To execute on strategies, managers must persuade and justify their decisions to employees, shareholders, and other stakeholders (Pfeffer 1981, Fiss and Zajac 2006, Helfat and Peteraf 2015). When managers can clearly explain the causal factors underpinning their decisions, it can enable stakeholders to see a logic behind decisions, thereby facilitating a shared vision, as well as their buy-in and support for execution (MacLennan and Markides 2021, Ehrig and Schmidt 2021). Moreover, well-articulated causal explanations provide a framework for managers to justify their decisions to external stakeholders such as investors, customers, and regulatory bodies. By presenting a causal logic for their strategies, managers can enhance the credibility and legitimacy of their actions. This kind of causal reasoning has been found to be particularly important for firm performance in environments characterized by uncertainty and complexity, where the ability to convincingly explain and justify decisions can be as crucial as the decisions themselves (Bettman and Weitz 1983, Pfeffer 1981). For example, even if a decision did not lead to superior performance, this may be attributed to a bad draw due to uncertainty, and shareholders can trust the decision-maker because they

perceived the causal explanation behind the decision to be sound.

Additionally, causal explanations can reflect decision-makers' theories about what drives value creation, which has received increasing attention in strategy research as a key determinant of better strategies (Felin and Zenger 2009, Csaszar 2018, Camuffo et al. 2020, Agarwal et al. 2023). Felin and Zenger (2017) highlight that decision-makers with a unique and clear theory grounded in a causal logic about what leads to superior performance – by providing a hypothesis on the assets, knowledge, and activities that enable a solution for a particular problem – can craft better strategies and superior firm performance. Similarly, Sørensen and Carroll (2021) argue that all great strategies are causal arguments with clear and logical reasoning. Effective strategic decision-making has been also posited to involve theorizing and explaining of the impact of key factors on performance, a process that resembles the elaboration and justification of initial beliefs through logical reasoning (Felin and Zenger 2009). This process of theorizing involves not only the identification of factors that influence performance, but also a detailed explanation of how these factors interact and lead to specific outcomes. Causal explanations can thus represent contextualized statements for a focal problem that can be derived from or embody the forward-looking theories that decision-makers hold.

2.2 The impact of machine predictions on causal explanations

The integration of machine predictions into strategic decision-making processes presents new challenges and opportunities. Agrawal et al. (2018) theorize that AI improves decisions by providing improved predictions, which are complementary to human judgment especially under conditions of uncertainty. Growing research provides supportive empirical evidence (Kleinberg et al. 2017, Hoffman et al. 2018), suggesting that the use of machine learning algorithms can significantly aid in decision tasks involving prediction. At the same time, recent work has highlighted that AI may not always improve decisions in organizations, as it introduces unique challenges such as biases embedded in data (Choudhury et al. 2020), opaqueness of AI "black boxes" (Lebovitz et al. 2022), and the need to consider broader sets of objectives or preferences beyond the algorithm's predictions (Kim et al. 2024). Recent meta-analytic evidence (Vaccaro et al. 2019) shows that human-AI combinations often perform worse than the best of humans or AI alone, with significant heterogeneity based on task type and AI performance relative to human performance.

Despite growing research on the impact of predictive algorithms on decisions, there has been less insight on whether and how machine predictions affect the way that decision-makers causally explain

and reason through their decisions. Understanding this question is important, because providing causal explanations has been posited as a key role that managers can play in a world of machines (Deming 2017, Beck and Libert 2017, Luo et al. 2021).

Theoretically, the effects of machine predictions on causal explanations are ambiguous. On the one hand, machine predictions have the potential to improve causal explanations. Building on Agrawal et al. (2019) who highlight the complementarity between predictions and judgment, Athey et al. (2020) argue that AI's augmentation of human activities can be profit-enhancing in complex decision areas, conditional on decision rights and aligned incentives. Moreover, by taking over repetitive tasks, AI can ease the multitasking problem and thus free decision-makers to focus on more cognitively complex tasks that require deeper thought, such as developing causal explanations. This re-allocation of effort can lead to more effective use of time and cognitive capacity (Simon 1947, Autor 2015, Agrawal et al. 2018, Camuffo et al. 2022). Grennan and Michaely (2020) similarly suggest that when stock analysts need to juggle multiple tasks, being exposed to AI advancements in their area of expertise redirects analysts' attention to tasks complementary to AI, such as engaging more with investors and working on more stocks, which raises the possibility that the use of machine predictions may also potentially improve decision-makers' causal explanations of their decisions by reallocating their efforts to reasoning and communication.

On the other hand, using machine predictions may also worsen decision-makers' causal explanations. Dell'Acqua (2022) suggests that decision-makers may become complacent and reduce their effort when working with algorithms, which may also potentially extend to reasoning about explanations, instead of reallocating their efforts to enhancing them. The division of labor induced by AI—by separating prediction from decision-making—may also lead to excessive modularization, which has been theorized to diminish decision-makers' understanding of complex systems (Siggelkow 2002, Ethiraj and Levinthal 2004, Billinger et al. 2014). This challenge may be particularly acute for inexperienced decision-makers who may not have developed as deep an understanding, especially in situations where causal explanations are inherently complex, leading them to be unable to develop clear causal models or integrate machine predictions into the causal model they have in mind.

In sum, the development of clear, insightful, and persuasive causal explanations plays a fundamental role in strategic decision-making, as it is crucial for managerial rationalizing and strategy execution in the face of uncertainty and complexity. Theoretically, using machine predictions as decision aids presents both opportunities and challenges in developing causal explanations, and this paper aims to explore these effects empirically as well as the mechanisms through which machine

predictions affect decision-makers' causal explanations.

3 Experimental context and setting

3.1 Why study mutual fund analysts?

To examine the impact of machine predictions on causal explanations, we partnered with a leading financial research firm headquartered in the United States to run an experiment across their mutual fund analysts.

Mutual fund analysts produce reports that provide recommendations for investors, with insights on the fund's strategy, holdings, risk-taking, and performance in the context of changing market conditions. Mutual funds are the predominant type of investment vehicle for U.S. households, with 52.3% of households (amounting to 68.7 million households) holding mutual fund investments in 2023. As of 2023, global mutual fund assets totaled approximately $65.5 trillion, with nearly 8,000 unique funds in the U.S. alone managing $18.9 trillion in assets (Board of Governors of the Federal Reserve System (US) 2023). Given the large number of mutual funds and the variety of investment strategies available on the market, investors seek analyst reports to better inform themselves about potential funds to invest in, as well as their broader investment strategy.

Analyst reports contain two key components. The first component is a rating for the mutual fund, which is the analyst's recommendation on whether to invest in the fund based on their forward-looking predictions of fund returns. This recommendation can be in the form of a binary recommendation, or as a rating that scales between "not recommend" to "highly recommend". Previous studies have shown that investors react to these mutual fund ratings, with higher-rated funds attracting significantly higher net cash inflow compared to lower-rated funds (Sirri and Tufano 1998, Gruber 1996, Goetzmann and Peles 1997), suggesting that the ratings decisions that fund analysts make have broader implications beyond their own company to investors in the market and the efficiency of the mutual fund industry.[3]

The second component of these reports is the analyst's causal explanation for their ratings decision. Analysts specialize in certain types of mutual funds and conduct in-depth quantitative and qualitative research: in addition to tracking changing market conditions, analysts review the

[3]However, this is unlikely to introduce the concern that analyst ratings will have a direct impact on fund returns within a short time frame. Fund managers construct portfolios of often a large number of assets. For funds that invest in fixed income assets, the returns of the underlying assets cannot be affected by fund ratings. For actively managed equity funds, a positive rating of the fund is unlikely to provide signals strong enough to affect the performance of individual underlying firms within the portfolio, especially within a 3-month period.

fund's investment strategy, documents issued by fund managers, as well as news articles that disclose information related to the fund or its parent firm. They also interview fund managers and other key decision-makers (e.g., managers at the parent firms, traders, and risk managers) to gather first-hand information, cross-validate secondary information, and discuss any concerns. They then piece together all of this research to provide a coherent reasoning to investors about what factors they believe causally drive fund performance and the underlying reasons why they assigned the rating they did for their forward-looking outlook on the fund.

These analyst explanations are generally valued more highly than the ratings themselves – by investors as well as the broader industry of financial research companies who employ analysts. Interviews with analysts as well as industry reports highlight that client demand for fund research is driven by the novel insights and "angles" or causal models they can glean from analyst reasoning to assess the fund, in hopes of better informing their investment strategies. Coherent causal models portrayed by the analysts help persuade investors to incorporate analyst research into their investment decisions. Studies have documented that analyst reports that provide better reasoning to investors gain more visibility and receive larger market reactions (Bradshaw et al. 2021, Campbell et al. 2019). The importance of causal explanations is also reflected in the way that financial research is sold to clients. While all clients can view the ratings, only clients with premium subscriptions have access to detailed analyst reports that offer causal explanations for these ratings. In fact, while ratings can be provided for a larger set of funds, only a small subset—those covered by analysts—include these causal explanations, leading companies to carefully select which funds are in their analyst coverage. This has implications for analyst incentives as well. While analysts generally do not have high-powered incentives, their career concerns are tied to the reputation they can establish from their reports, which can provide them with better outside options (Cornaggia et al. 2016, Kempf 2020, Lourie 2019). Star analysts may not always produce ratings that accurately predict future outcomes, which represent a random draw from a distribution with high variance. However, analysts can demonstrate their expertise and research capabilities by presenting and explaining their causal models persuasively in their reports, which can account for key sources of uncertainty regardless of what future outcome is realized.

Recent years have seen the growth of the global fund market, which has increased the sheer number of mutual funds and subsequently the demand for fund analyst research. The total number of global open-ended funds increased from 66,362 to 137,892 between 2007 and 2022, making it impossible for research firms to cover the entire universe of funds. For example, the financial research

company that we collaborated with for our experiment—one of the largest in the market—was only able to cover about 10% of the mutual funds with their analyst research.

This has underscored the demand for analyst reports and the explanations contained in them, as evidenced by qualitative interviews we conducted with financial advisors and wealth managers – the primary consumers of fund analyst reports. They highlight how customers of these reports use and actively integrate analyst explanations into their investment decision-making process, portfolio construction activities, and client communication. The quality and depth of the analyst's explanations directly influence their willingness to recommend funds to clients, with practical implications for real investment outcomes. Financial advisors state that analyst explanations are an "integral part of [their] decision making process," and highlight that they especially value analysts' explanations beyond facts and data points, which directly inform their evaluation framework:

> "These recommendations aren't just purely based on on performance, [...] there are other reasons, too. Obviously, that's why I'm looking at for their insights."
>
> "I'm looking more for what's going on under the hood of the fund."
>
> "We've seen it again and again. [Fund managers] go through periods where they could do no wrong, and then they hit periods where they can't get anything right. So are they lucky, or are they skilled?"

In addition to supporting their investment decision-making process, analyst explanations are used to communicate with their clients. Financial advisors emphasize that they can help explain and justify investment decisions to clients, articulating that the value of third-party validation is crucial for client conversations.

> "It helps with investment decisions. If clients have a question, I will scan it to them, or we'll sit down and go over the report. ... It's nice to have a 3rd party. I have my biases. You have your biases. We all see the world differently."

This increasing demand from clients, as well as the growing availability of data, has accelerated the adoption of AI-related technologies in the financial research industry to help reallocate analyst effort from prediction to explanations. Both incumbents and startups alike have invested in developing algorithms to perform analytical tasks that historically relied on financial experts, exploring how algorithms can help evaluate financial securities based on big data analytics, leveraging available quantitative data also from alternative sources (Bollaert et al. 2021, Grennan and Michaely 2020). Large banks, including JP Morgan, Citigroup, and Deutsche Bank, were reported to have hired

thousands of AI-related roles in 2023 to use AI to model risk and conduct data analytics (Shaw and Gani 2023).

This context thus provided a unique opportunity to understand how the use of predictive algorithms might affect decisions and causal explanations. The nature of this work shares many key features with other kinds of complex knowledge work and strategic decision-making, where decisions are forward-looking, complex, and made under uncertainty, with both a predictive component and an explanation component to explain the causal reasoning—making it plausible that insights from our study could potentially generalize to other similar settings that have often been challenging to study empirically due to small sample or measurement issues.

3.2 Experimental setting and data

We collaborated with a leading company in the financial research industry to run an experiment across all decisions made by their full set of mutual fund analysts globally (encompassing the US, EMEA (Europe, the Middle East, and Africa), Asia, and Australia). It was at the forefront of adopting AI-driven technologies to aid the process of financial research, having been one of the first companies to launch financial research products leveraging predictive algorithms.

At the time of the experiment, the company had developed a machine learning algorithm to help analysts make decisions on mutual funds, but had not yet rolled them out across analysts. The algorithm was a tree-based model trained on analysts' past decisions and all quantifiable data that analysts reviewed to make their decisions, which included all available data except private information collected by analysts, such as interviews of managers and any tacit knowledge analysts accumulated from experience (e.g., the style of fund managers and who may be prone to over-optimism). The firm had closely monitored the algorithm's performance since its development and reported that the algorithm clearly separated over- and under-performing funds into categories following the company's rating methodology: highly-recommended funds outperformed their category average by about 1%, and not-recommended funds underperformed their peers by 1% over a 12-month period following the rating. After a few years of observing the performance of algorithmic ratings, piloting changes, and back-testing them to ensure their high level of quality, the firm decided to introduce these machine predictions as a decision aid for analysts. While analysts had not previously used predictions from these algorithms, they were continuously updated on their development and informed of their methodology and overall quality.[4]

[4] Analysts were continuously updated on these algorithms over several years, so we do not expect any anticipatory

We collaborated with the firm at this juncture to randomize the provision of machine predictions across analysts, which provided a unique opportunity to measure how analysts' decisions and causal explanations changed when working with machine predictions.

Decisions: we collected all analysts' fund rating decisions between April 6, 2021 and January 19, 2022, which covered 1,780 rating decisions in total. Mutual fund analysts at the firm specialized in certain asset classes (e.g., equity, fixed income), but covered funds both inside and outside of their specialization, which was pre-determined by managers and could not be changed spontaneously by analysts. For each fund in their coverage, analysts updated their decisions on the fund's ratings at least once every year. Because these ratings were forward-looking in nature, we also collected data on these funds' returns in the subsequent three months and six months, respectively, which enabled us to assess these decisions by evaluating the extent to which funds recommended by analysts provided higher risk-adjusted returns on average in the months that followed.

Causal Explanations: we collected a measure of the quality of analysts' causal explanations between April 6, 2021 and January 1, 2022.[5] Explanations for decisions are rarely articulated or observable in full in practice, and generally difficult to assess according to stakeholders who would be relevant for the decision. In this company, all analyst ratings and reports had to be approved by an internal expert committee before they could be released publicly, which provided us with a unique opportunity to collect a measure of explanation quality. All committee members were blind to the experiment and assignment to treatment, and the evaluation criteria used by the committees stayed constant throughout this period without any changes. The committees evaluated analysts' causal explanations for each of their fund rating decisions by answering the following questions on a 1 to 10 scale: (1) How would you grade the analyst's knowledge of the strategy? (2) How would you grade the analyst's reasoning behind the ratings? Committees were composed of one to four experienced analysts who managed these teams and held director-level positions. Single member committees were the most common, evaluating 78.48% of all rating proposals. Committees with two, three, or four members evaluated 15.79%, 5.56% and 0.16% proposals, respectively. The composition of the committee members varied from fund to fund, and was constructed based on committee members' expertise and availability. For each fund rating submitted by an analyst, we standardized evaluator scores and averaged them across the committee. Appendix Section A.1 provides two examples of

effects in the months during the experiment.

[5] On January 1, 2022, the company decided to remove the requirement for internal review committees to provide explicit scores in order to reduce their workload, which meant that they only provided approval decisions and feedback comments back to analysts from this date onwards. This provided us with 1,313 reasoning scores from committees.

these reasoning scores accompanied by the summary section of the corresponding analyst reports, contrasting a report that received a high reasoning score (example 1), with another that received a low score (example 2).

In addition to these reasoning scores, we collected text and survey data to capture richer measures of analysts' causal explanations. We use the full text of analyst reports, which provide us with direct insight into how their causal explanations changed. Furthermore, when analysts submitted each rating decision and report, they listed the causal drivers that drove their decisions in a survey built into the workflow, which allowed us to capture the number of key factors they proposed as driving their decision. We also leveraged large language models (LLMs) to extract causal statements from the text of analysts' reports. We conducted in-context learning in GPT-4 to identify causal statements across the reports, decomposing each extracted statement into its cause and effect (see Appendix II.2 for examples of the causal statements extracted). In addition to these measures, we directly analyzed the text that analysts wrote in their reports, using various natural language processing techniques to evaluate the length, readability, and coherence of the text. We used two widely used readability measures, Flesch-Kincaid and Coleman-Liau, both of which provide a decreasing measure of readability (the higher the score, the more difficult it is to read, associated with a higher grade level of education). The Flesch-Kincaid grade-level score factors in both the number of words in sentences and the difficulty of words based on the average number of syllables per word (Kincaid et al. 1975). The Coleman-Liau index provides an alternative measure based on sentence length and the number of characters per word to approximate readability (Coleman and Liau 1975). Finally, we computed a coherence score based on the coherence of topics discussed in the reports using topic modeling (Röder et al. 2015).

Perceptions of machine predictions: we collected data on analysts' perceptions of machine predictions for those assigned to treatment. For each fund, analysts assigned to treatment rated on a scale of 1-10 how accurate they found the machine prediction to be and the extent to which it contributed to their decision.

We observed each of these measures across all decision-makers and decision types across the company, which spanned a range of analyst experience and decision complexity. The tenure of analysts in the company ranged from 0 to 29 years, with approximately 25% of analysts with more than 10 years of experience in the firm. An important distinction in the company was whether analysts were new to the firm, as the company had a unique methodology and new analysts had to undergo substantial training to learn how to evaluate mutual funds according to this methodology.

The type of decisions analysts made also varied in terms of their complexity based on the fund type, providing variation on decision complexity, a key attribute of strategic decisions (Csaszar 2018). For each fund, analysts assessed the quality of the fund's management team and the fund's investment strategy. Evaluating the team behind the fund included assessing elements such as the ability and experience of the managers, the availability of supporting resources, conflicts of interests, risk-taking tendencies, and fit with the investment strategy. Evaluating the fund's investment strategy involved assessing whether the fund had distinct investment processes that were clearly defined and consistent with the investment process and performance objective, and that produced standout results in a systematic and sustainable way.

This decision process was typically more complex for active equity funds compared to all other fund types, a key distinction made by the company. Active equity (AE) funds are portfolios of stocks actively managed by fund managers, meaning that the fund's managers and investment strategy change more substantially and frequently, the returns tend to be more volatile, and there is a lot of information and news coverage about the stocks in the portfolio that drives these changes. Non-active equity (NAE) funds include passive equity funds, fixed income funds, allocation and alternative funds. These funds change less over time in terms of their management and strategy attributes, and thus generally involve assessing how the market has changed since the last rating update and what this might mean for the performance of the fund. AE funds require extensive analysis for each rating update, to understand and evaluate changes in the team, investment philosophy, strategy, risk-taking, resources, and execution. Analysts also analyze the portfolio holdings of funds using extensive databases and analytical tools to evaluate each of these changes relative to other similar funds and how they might perform in the current market environment. Decisions on AE funds thus tended to be more complex compared to NAE funds, which we refer to as "complex" (AE) versus "simpler" (NAE) funds henceforth in the text. Importantly, although analysts have a primary fund type that they specialize in, they cover both simpler and complex funds due to capacity and legacy reasons.

These differences in the complexity of analysis do not indicate that ratings decisions or causal explanations on simpler NAE funds are less important than complex AE funds for the company or its client investors. Passive investment has been on the rise in the past decades due to their relative low cost and higher risk-adjusted returns compared with active investment on average. Assets under management by passively managed US mutual funds and exchange traded funds amount to $13 trillion, which overtook their actively managed counterparts for the first time in December 2023

(Schmitt 2024). The simpler NAE funds that analysts were assigned to cover were thus carefully selected to help guide investors in navigating their choice of investment vehicles among these growing set of choices.

3.3 Descriptive analysis on causal explanations

At baseline, causal explanations that received higher reasoning scores identified fewer key causal drivers behind the rating to explain and justify the decision (Figure A.1 (b)). Because possible drivers are numerous, spanning more than fifty high-level categories, the primary goal of the explanation in the analyst report was to pare down the list of drivers to identify ones that mattered most, which could take considerable effort. We observe this pattern across both simpler and complex funds (Figure A.3). On average, experienced analysts received higher reasoning scores on their causal explanations (Figure A.1 (a)).

We also find that better decisions and causal explanations show a positive correlation at baseline. Recommended funds accompanied by explanations with lower reasoning scores were less likely to observe higher risk-adjusted returns compared to non-recommended funds, whereas recommended funds with higher reasoning scores were more likely to observe at least as high returns compared to non-recommended funds (Figure A.1 (d)) – suggesting a positive association between the performance of ratings decisions and reasoning scores.

We also observe that conditional on being recommended, ratings with causal explanations that received higher reasoning scores were marginally more likely to induce higher fund inflows from investors one month after the publication of analyst reports (Figure A.15(a)), suggesting that these reports were more persuasive. Similarly, conditional on being not recommended, ratings with causal explanations that received higher reasoning scores induce higher fund outflows from investors ((Figure A.15(b)).

3.3.1 Robustness check on reasoning scores

We also conducted a mixed-method evaluation consisting of an experiment and qualitative interviews to validate our internal expert committee's reasoning scores. This involved 16 professional financial advisors (median age 59 years, 75% with over 21 years of tenure) who collectively managed approximately $4.8 billion in assets – representing the firm's most sophisticated institutional investor client base, recruited through our partner company's consumer service vendor.

Participants each evaluated four randomly drawn analyst reports from a pool of 20 analyst

reports on US equity funds that all received the same analyst rating, suggesting that the analysts highly recommended these funds to investors (see Appendix Section A.2 for full details). Controlling for the rating decision, these reports varied in the reasoning scores assigned by the internal expert committee, and asked participants to assess the analyst's explanations on dimensions including usefulness, clarity, demonstrated knowledge, and logical reasoning, before stating their likelihood to invest in the fund recommended by analysts. The fund names and identifiers were masked so that participants could not search for more information about the funds. Importantly, participants were not provided with the internal expert committee's reasoning scores for the reports.

Following the evaluation, advisors participated in semi-structured 30-minute interviews exploring how they value and use analyst explanations in practice. Figure 1 and Table 1 show that external experts and internal expert committees are highly aligned with how they evaluate analyst explanations along all dimensions. "Low quality explanations" indicate reports that received below-median reasoning scores from the internal committee; "high quality explanations" indicate those that received above-median reasoning scores. We observe that external financial advisors are significantly more likely to find reports that received higher scores from the expert internal committee to be more knowledgeable, logical, useful, and clear, and indicate a higher likelihood of investing in these funds.

Furthermore, qualitative interviews following the experiment emphasized the importance of analyst knowledge and causal reasoning in these reports – the two dimensions assessed by the internal expert committee to construct reasoning scores. One advisor highlighted,

> "There's a lot of opinions in this thing and that's great. I don't mind opinions. I just don't want those opinions to be oxymorons or just filling air."

Another underscored the importance of a clear and well-reasoned argument that helped them understand the logical progression from data points to conclusions.

> "I like more of a logical layout with a progression of flow and justification for choices made by the manager. ... More information, but condensed in a more concise way and presented in a logical flow."

Specifically, advisors highlighted that they sought analyst explanations to provide insight into which key factors matter most and the reasoning behind their prioritization, in line with the patterns we observe that analyst reports with higher reasoning scores report fewer causal drivers.

> "Are they all equally weighted? Are they all equally important? I kind of feel like maybe momentum isn't as important if the rest of things are through the roof. Are these (factors) weighted differently? And if so, I'd like to know which is the least important. Say, with values the most important, or qualities most important and low volatility being the least important."

Together, these results provide additional support that validates the generalizability of the internal review committees' evaluation of analyst explanations.

4 The machine prediction experiment

4.1 Experimental design

The experiment ran across all 97 mutual fund analysts at the company globally. We randomly assigned half of the analysts (48) to treatment, which provided them with machine predictions on ratings. Half of this treatment group additionally received a figure that provided some algorithm explainability. We implemented the experiment for 6 months from July 18, 2021 to January 19, 2022 and observed a total of 1,780 decisions.

Before the experiment, analysts made decisions without any algorithmic aids. They conducted research using an internal corporate web interface, which displayed information relevant to fund analysis and provided links to other databases (Figure A.2). The top of this interface highlighted the most recent rating that was assigned for that fund, and several tabs of subpages provided data on the fund's historical performance, fee structure, and details about the fund management team and the parent firm. Analysts analyzed this data along with others that they collected on their own (e.g., via interviews) to assess the fund strategy and its potential to outperform its peers. They then synthesized this analysis to rate both the team of fund managers and the process/strategy of the fund, which were then aggregated along with other metrics to classify the fund into one of five categories displayed to investors indicating its investment quality – the top three of which were considered to be "recommended", and the bottom two which were not recommended.[6] The final aggregate rating constituted whether this fund was shown to carry their recommendation. Analysts then elaborated on their reasoning in a report that provided an explanation with relevant analysis on reasons driving the rating decision. These were all submitted via the corporate interface, which were then reviewed by an internal expert review committee, who was blind to the experiment as well as analyst assignments to treatment. The composition of the committees varied from fund to

[6] Analysts observed an aggregated rating and adjusted their sub-ratings accordingly until a final decision is made.

fund, and was constructed based on committee members' expertise and availability.

Within this context, analysts were randomly assigned to one of three experimental groups: (1) a control group, who made decisions according to business-as-usual; (2) a treatment group who received machine predictions on ratings; and (3) a treatment group who received machine predictions on ratings with some form of algorithm explainability. The control group saw no changes on the interface and made decisions as usual (Figure 2 (a)). Analysts assigned to either treatment group were shown an additional line of information at the top of the interface that provided a machine prediction for the fund's current rating, which was displayed above the last assigned rating for the fund (Figure 2 (b)). Treatment 1 provided machine prediction only, while Treatment 2 additionally provided a figure that highlighted the ten key features that contributed most to the algorithm's prediction for that fund (Figure 2 (c)).[7] This explanation was based on Shapley values, a concept developed in the computer science literature as one method for "explainable AI", using cooperative game theory to determine the average marginal contribution of each feature value across all possible values in the feature space (Lundberg and Lee 2017). These treatments were extensively tested and piloted with a small group of key stakeholders at the company, which included the team that managed the interface, the team that developed the algorithm, the head of investment analysis, and directors of research teams who managed analysts and previously were senior analysts themselves.

Half of the analysts were randomly assigned to control, and the other half to one of the two treatment conditions, using stratified randomization. We stratified on the fund type that they specialized in and analyst tenure. The randomization was coded into the backend of the corporate interface, which enabled us to estimate intent-to-treat effects. Analysts rated and published a total of 1,137 funds during the experimental period, averaging approximately 2 funds per analyst per month on average.

As analysts covered different funds individually, the nature of their work did not involve collaboration, and all analysts worked independently. We did not expect non-compliance, as random assignment was coded in the backend of the corporate web interface, and fully controlled what was shown on the interface for the analyst. However, it is possible that analysts may not pay attention to the treatment intervention, which would bias us against finding a treatment effect. The experiment was implemented during the Covid-19 pandemic, when all analysts worked from home and individually accessed the corporate interface – further limiting the likelihood that other analysts

[7] These features were common fund attributes that all analysts were familiar with, such as the fund's fees, its past performance, and fund age.

could observe different versions of the interface. The interface was also actively undergoing development during this time, which meant that analysts were aware that certain features were becoming available at different times across different funds. Given these contextual details, we and the key managers at the company did not expect substantial spillover effects across analysts to drive potential treatment effects. Since the machine predictions are specific to each fund, even if analysts in the control learned about others in the treatment group, they would not have access to the machine predictions of the funds they were assigned to work on.

As expected by randomization, experimental groups were well-balanced across baseline indicators (Table 2). We observed no attrition among analysts, and recorded the published rating decisions of all 1,780 funds. However, the number of observations varied across outcomes for two reasons. First, the company stopped collecting all reasoning scores starting January 2022 – resulting in a total number of observations of 1,313 for reasoning scores. Second, for fund returns, the sample size was reduced from 1,780 to 1,668 because of missing data on risk-adjusted fund returns. This is because funds lacking complete 36-month return histories cannot have their risk-adjusted returns estimated. This should not bias our treatment effects, since for reasoning scores it means that we simply collected data on both groups for a shorter number of months, and for fund returns, the availability of prior 36-month data should not correlate with treatment given the randomization. Indeed, the share or number of missing data for any of these outcomes does not vary between experimental groups (Table 3).

4.2 Econometric specification

To evaluate how machine predictions affect analyst decisions and causal explanations, we run the following difference-in-difference specification as our base model as pre-registered:

$$y_{ift} = \alpha Treat_{ift} + \beta Post_{ift} * Treat_{ift} + \gamma Post_{ift} + \delta_j + \eta_t + \epsilon_{ift} \tag{1}$$

y_{ift} is the outcome of interest for fund f evaluated by analyst i at month t, which includes: (1) a binary indicator of whether the analyst decided to change the fund's currently assigned rating; (2) the reasoning score for the analyst's causal explanation for the ratings decision as assessed by the internal expert committee; (3) the number of self-reported decision drivers that drove the analyst's decision; (4) the number of causal statements identified in the analyst report. $Treat_{ift}$ is an indicator variable that takes value 1 for analysts assigned to treatment (pooling across both treatment groups)

and 0 otherwise. $Post_{ift}$ is an indicator variable that equals 1 for the experimental period and 0 otherwise. While fixed effects are not necessary for identification given that treatment is randomly assigned, we run this specification with randomization strata (δ_j) and month (η_t) fixed effects to account for any random differences across experimental groups and soak up noise.[8] We cluster standard errors at the analyst level, which is the unit of randomization.We additionally estimate p-values based on randomization inference to all main regression analysis. The randomization inference approach, which does not rely on asymptotic assumptions, is particularly well-suited for analyzing data from randomized experiments with small samples (Athey and Imbens 2017). Our findings remain qualitatively unchanged and fully consistent when applying the randomization inference approach.

β identifies the treatment effect of providing machine predictions, which is the main coefficient of interest. We also use the same specification with separate indicators for each treatment group to compare the two treatments and assess the impact of providing explainability for algorithmic predictions. The key identifying assumption is that analysts assigned to treatment did not have systematically different trajectories of rating decisions and reasoning compared to those in the control group for reasons other than the algorithmic prediction treatment, which was randomized.

We then evaluate whether machine predictions led to better decisions by assessing the extent to which analysts' recommended funds observed higher risk-adjusted returns in the months following their ratings decision. We evaluate fund returns over a three- and six-month period, using a monthly time series of fund return data from Morningstar Direct and the Center for Research in Security Prices (CRSP) Mutual Fund Database. Following the accounting and finance literature on ratings and fund returns, we calculate risk-adjusted returns to account for variation in risk-taking across funds, using the benchmark Fama and French (2015) five-factor model. We estimate factor loadings by running 36-month rolling regressions where each fund's excess returns over the prior 36 months are regressed on the relevant factor returns. The factor loadings are then used to adjust the raw monthly return data to estimate risk-adjusted returns. We also run the Fama and French (1993) three-factor model and the Carhart (1997) four-factor model to check the robustness of our results, which provide consistent results across risk-adjustment methods.

[8]Results using alternative model specification using analyst fixed effects instead of strata fixed effects are reported in Appendix Table A.34, Table A.35, Table A.36, and Table A.37 to assess whether the observed effects are driven by within-analyst variation. The inclusion or exclusion of analyst fixed effects does not affect identification, given that the randomized assignment is at the analyst level. The results remain fully consistent with the main model specification with strata fixed effects. The estimates for reasoning scores are noisier, likely due to the smaller sample size for this outcome. When treatment is randomly assigned with strata, strata fixed effects generally yield the most statistically efficient estimates (Imbens and Rubin 2015).

$$
\begin{aligned}
r_{ift} = \alpha_1 Treat_{ift} + \gamma Post_{ift} + \nu Recommended_{ift} \\
+\beta_1 Treat_{ift} * Post_{ift} + \beta_2 Treat_{ift} * Recommended_{ift} + \beta_3 Post_{ift} * Recommended_{ift} \\
+\beta_4 Treat_{ift} * Post_{ift} * Recommended_{ift} \\
+\rho * FundAge_{ft} + \sigma * FundSize_{ft} + \delta_j + \eta_t + \epsilon_{ift}
\end{aligned} \tag{2}
$$

We regress these risk-adjusted fund returns on a variation of specification (1) above, which includes interactions with a binary indicator of whether a fund was recommended by the analyst. $Recommended_{ift}$ is an indicator variable that takes value 1 when a fund is recommended by the analyst and 0 otherwise. Following best practices in the finance and accounting literature (Armstrong et al. 2019), we also control for key fund-level characteristics, fund age and fund size (proxied by net assets). β_4 identifies whether recommended funds by analysts randomly assigned to treatment observed higher fund returns compared to those recommended by analysts assigned to the control group.

In addition to average effects, we explore heterogeneity in treatment effects for all outcomes by interacting binary indicators of pre-registered variables. We focus on two key dimensions of heterogeneity: fund type, which determines the complexity of decision-making that analysts engage in, and analyst experience, which is associated with domain expertise. For fund type, we use a binary indicator for complex funds, which represent 45.35% of all fund decisions during the experimental period. For analyst experience, we use a binary indicator that equals 1 for new analysts with less than three years of experience, and 0 otherwise. We use this cutoff as analysts were promoted to an official "analyst" position from the associate level within three years on average.[9]

[9]We check the robustness of our analysis with experience as a continuous measure and reported these results in Appendix Table A.42 and Table A.43. We find fully consistent results as in our main tables (Table 4 and Table 5).

5 The impact of machine predictions on decisions and causal explanations

5.1 The impact of machine predictions on decisions

To assess whether analysts considered machine predictions in their decisions, we first examine whether treatment impacted analysts' likelihood of changing the rating assigned to the fund.[10] This analysis serves as a manipulation check for us to know that treatment indeed had an impact on analysts' rating behaviors: if they are more or less likely to change the fund rating as a result of being randomly assigned to receive machine predictions, this suggests that the treatment had some effect on analyst behavior. Table 4 Panel A Columns 1-4 show estimates of the average treatment effect across different specifications, with Column 4 showing that analysts randomly assigned to receive machine predictions were on average 8 percentage points more likely to change the fund rating (p-value=0.01), a 71% increase over 11% of rating decisions that were changed in the control group at baseline. This suggests that machine predictions had a fairly large impact on analyst decisions, as they induced a change in the assigned fund rating. The estimated treatment effect does not vary significantly across decision type or analyst experience (Table 4 Panel A Columns 5-6), suggesting that analysts largely responded to the machine prediction treatment. This treatment impact is unlikely to be driven solely by the fact that analysts receive a rating to anchor on, because all analysts in both control and treatment groups were provided with the live analyst rating for the fund that they assessed, which represents an analyst-generated prediction for all funds.

We then explore the extent to which machine predictions improved decisions, by evaluating the risk-adjusted returns of analysts' recommended funds in the months that followed. To analyze whether treated analysts' ratings performed better, we compare whether the risk-adjusted abnormal returns of funds that treated analysts recommended were higher compared to those of the control group.[11] Column 1 of Table 4 Panel B shows that on average, funds recommended by treated analysts outperformed those of analysts assigned to the control group by 0.89 percentage points (89

[10] The live rating is generally assigned one year previously. It is worth noting that changing the live rating has also been seen as requiring more confidence: analysts are more likely to maintain the status quo unless they have enough confidence that this update is correct (Fernandez and Rodrik 1991, Samuelson and Zeckhauser 1988).

[11] We follow the finance literature in analyzing risk-adjusted abnormal returns, which allows us to determine the performance of a financial asset or portfolio when compared to the overall market or a benchmark index. It has been used to help identify an investor's skill on a risk-adjusted basis, finding that high-skill investors consistently generate positive risk-adjusted returns (Soydemir et al. 2014, Bach et al. 2020). Studies in accounting and finance have also analyzed risk-adjusted returns to evaluate the performance of mutual fund ratings (Armstrong et al. 2019, Ertugrul and Hegde 2009, Luo et al. 2015), as ratings that better predict future fund returns generate more profits for investors who build their portfolios following these ratings.

basis points), but this estimate is not significant (p-value=0.77). On average, funds recommended by analysts assigned to the control group observed returns of -4.5%.[12]

These average effects mask some heterogeneity in treatment effects, which appear to vary depending on the type of decision involved – improving simpler decisions while worsening more complex ones. Simpler funds recommended by treated analysts observed 9.9 percentage point higher returns (p-value < 0.001) compared to those in the control group – suggesting that these simpler decisions improved as a result of using machine predictions. In comparison, the treatment effect estimate on the returns of more complex funds recommended by analysts was 16.8 percentage point lower (p-value=0.002).

Moreover, the treatment effect appears to be driven by new analysts with less experience. For those assigned to treatment, simpler funds recommended by new analysts observed 11.6 percentage point higher returns (p-value = 0.002) compared to those in the control group, and more complex funds recommended by new analysts observed 16.8 percentage point lower returns comparatively (p-value=0.04). Treatment effects are more attenuated and less precisely estimated for experienced analysts for both simpler and more complex decisions, with interaction coefficients in the opposite direction compared to new analysts for each decision type.

These treatment effects remain qualitatively robust when analyzing fund returns within 6 months, though attenuated (Table 4 Panel B Columns 4-6). The attenuated fund returns result within 6 months do not necessarily suggest that the treatment effect dissipates over time. Fund returns become more volatile and subject to a larger number of economy-wide factors in the longer run and become more challenging to predict, which is partly why investment research firms often seek to update ratings more frequently.

These treatment effects remain qualitatively robust with alternative model specifications that include analyst fixed effects (Appendix Table A.34), and split sample analysis (Appendix Table A.46-A.49).

5.2 The impact of machine predictions on causal explanations

We perform textual analysis on analyst reports to assess whether machine predictions affect how treated analysts write their reports. Similarly to our "Change Rating" measure, we trained a Light Gradient Boosting Machine (LightGBM) model with TF-IDF vectorization to evaluate whether the

[12]This is not atypical: the literature documents negative fund risk-adjusted returns on average, net of expenses and trading costs (Barras et al. 2010, Carhart 1997, Jensen 1968).

treatment condition to which an analyst was assigned could be predicted from the analysts' writing alone. If so, this would provide additional evidence that the machine predictions changed the way analysts make decisions and provide causal explanations. Our analysis indeed reveals that reports written by analysts in treatment and control groups are significantly different in their linguistics. Figure 3 shows that we can classify reports into treatment conditions with an accuracy rate of 85% using a standard TF-IDF model, suggesting that analysts write differently when they have been assigned to receive machine predictions. This classification accuracy is particularly high for reports written by new analysts evaluating complex funds, reaching an accuracy rate of 90%.

We next evaluate how machine predictions affected analysts' causal explanations, and find treatment effects to be in the opposite direction from decisions. Figure 4(a) and Table 5 Column 1 show that average treatment effects on reasoning scores are not economically or statistically significant (Kolmogorov-Smirnov test p-value=0.336; p-value=0.94). Confidence intervals on these estimates are quite wide, ranging from -0.3 to 0.3 standard deviations in either direction, making it difficult to rule out large effects.

We again find that treatment effects vary by decision type and analyst experience. However, in contrast to the treatment effect on decisions, machine predictions slightly worsened the reasoning scores of analysts' causal explanations for simpler decisions, while improving their explanations for complex decisions (Figure 4(b) and (c)). Treatment appears to have shifted the distribution of scores in both cases (Kolmogorov-Smirnov test p-values: 0.002 and <0.001, respectively). On average, differences are less precise (Table 5 Column 2) but similar directionally: algorithms deteriorated the reasoning for simpler decisions – which improved – by 0.12 standard deviations (p-value=0.6), while improving the reasoning of more complex decisions – which relatively worsened by 0.3 standard deviations (p-value=0.4).

These treatment effects are more precise when further separated by analyst experience, suggesting that new analysts were again more likely to be impacted. Figure 4(d)-(g) shows that treatment effects appear to be driven most by new analysts, whose reasoning for simpler decisions worsened, while it improved for complex decisions (Kolmogorov-Smirnov test p-values = 0.015 and < 0.001 for new analysts; 0.09 and 0.572 for experienced analysts, respectively). Similarly, Table 5 Columns 3-4 show that causal explanations of new treated analysts received approximately 0.7 standard deviations lower reasoning scores on simpler decisions relative to control analysts in the experimental period (p-value=0.06). In contrast, new analysts were more likely to see their reasoning scores improve on more complex funds, observing a 0.9 standard deviation higher treatment effect (p-

value=0.02) – even though they were more likely to make worse rating decisions. For both decision types, more experienced analysts were again less impacted, with attenuated coefficients in the opposite direction. While we observe a weaker impact of machine predictions on experienced analysts, it is also possible that reasoning scores for experienced analysts varied less if review committees were less sensitive to differences in explanations by experienced analysts given their reputation—which we cannot fully tease apart empirically. Columns 4 and 5 of Table 5 include committee fixed effects, with Column 5 restricting to reasoning scores by single member committees, which approved the majority of ratings during this period, thus providing more precision when including committee fixed effects.[13]

In addition to reasoning scores, we analyze the text of analysts' reports to better understand how their causal explanations change (Table A.2).[14] We use natural language processing techniques to evaluate the readability and coherence of the text, and find directionally similar effects as reasoning scores, though less precise. Treated new analysts produced causal explanations with lower readability and coherence on simpler funds, while showing higher readability and coherence on more complex funds, suggesting that they used more precise technical language and showed higher consistency in logic, in line with their higher reasoning scores.

5.3 The impact of algorithm explainability

Finally, we evaluate how algorithm explainability affected analysts' decisions and causal explanations, and observe limited differential effects (Appendix Figure A.8, Table 6). Across all outcomes, estimates for the explainability treatment are generally qualitatively similar to those for the treatment without explainability, across both average and heterogeneous effects.

However, Figure A.12(c) suggests that explainability generally increased the perceived contribution of machine predictions. Moreover, explainability increased reliance on machine predictions especially among experienced analysts, who generally tended to rely less on machine predictions overall, consistent with prior studies (Allen and Choudhury 2022, Kim et al. 2024).[15]

It is important to note that our results may reflect the limitations of the method chosen by the firm rather than the overall ineffectiveness of explainability. The Shapley values (a feature

[13]Committees with a single reviewer evaluated 20.6 ratings on average, while committees with two reviewers evaluated an average of 3.7 ratings, and committees with three and four reviewers evaluated 1.4 and 1 ratings on average, respectively.

[14]This current draft includes results on all published reports for U.S.-based funds.

[15]It is worth noting that our broader findings are also in line with recent studies that challenge the benefits of creating explanations for black-box models (Vaccaro et al. 2019), and raise concerns that such practices may in some cases do more harm than good (Rudin 2019).

attribution method often referred to as SHAP) used in our setting are one of the most widely adopted methods at the frontier of research on explainable AI so far by scholars and engineers in practice (Lundberg and Lee 2017), which seek to provide a practical way to offer goal-oriented, teleological explanations to inherently uninterpretable algorithms (Tomaino et al. 2020). More importantly, the issue may be the lack of true explainability in Shapley values, especially given recent advances in AI.[16] Numerous alternative explainability methods exist, such as rule-based surrogates and counterfactual explanations, each with distinct strengths and limitations (Burkart and Huber 2021). Alternative explainability methods should be investigated in future studies to understand how they may bridge the tension between algorithmic optimization based on correlations and human needs for causal understanding.

Together, our results suggest that while algorithms can help improve the decisions that they aid, they may deteriorate the causal explanation behind the decision and possibly introduce a tradeoff between the two. The findings raise the possibility that organizations may face a key challenge as they use algorithms to support managerial decisions, especially for less experienced managers. For those who make simpler decisions, algorithms may worsen their reasoning even when they improve the decision, while for those who make more complex decisions, algorithms may worsen decisions even while improving the reasoning behind them. The findings suggest that even when explanations are provided, the fundamental cognitive challenge remains.

6 Mechanisms

Our results reveal a striking tradeoff: machine predictions improve decision quality but degrade explanation quality for simpler decisions, while improving explanations but worsening decisions for complex ones. What might explain this pattern?

We propose that this tradeoff may reflect a cognitive integration challenge, in which analysts must reconcile machine predictions—often opaque and unfamiliar—with their own mental models to construct coherent explanations. When presented with machine predictions, analysts face a fundamental dilemma: either incorporate a forecast they do not fully understand and construct a post hoc explanation around it, or rely on their own forecast and causal model, which they can

[16] Machine learning algorithms are black boxes that are difficult, if not impossible, to fully explain. The tension between how algorithms operate and how humans make decisions is well-established in the explainable AI literature (Burkart and Huber 2021): algorithms can only optimize based on correlations, while humans need to understand and explain decisions by discovering causal dependencies. This gap becomes particularly problematic given the value and importance of causal explanations in our research setting.

explain. When machine predictions outperform analysts' own forecasts, this dilemma produces a tradeoff: incorporating the machine prediction improves decision quality but weakens causal reasoning; relying on one's own model results in worse decisions but more coherent explanations. If analysts are more likely to adopt machine predictions for simpler decisions, then this would help explain our results: relying on machine predictions improves these simpler decisions while worsening their explanations, and have the opposite effect for more complex decisions, where analysts rely on their own predictions and models.

We find some evidence consistent with this explanation. Survey questions built into the analysts' workflow provide insight into their perceptions of machine predictions and how they contributed to their decisions across decision type and analyst tenure. We observe that treated analysts reported greater reliance on machine predictions for simpler decisions than for complex ones (Figure A.12(a)). The standardized mean perceived contribution of machine predictions was 0.14 for simpler decisions (95% CI: [0.02, 0.26]) and -0.19 for complex decisions (95% CI: [-0.32, -0.07]), a statistically significant difference. This suggests that explanation quality deteriorates when analysts rely more heavily on machine predictions.

Moreover, this mechanism suggests that new analysts should struggle more with developing post hoc causal explanations when relying on machine predictions, as they lack experience in both forecasting and constructing causal narratives. Their inexperience is likely to make it more difficult for them to integrate multiple perspectives to develop a post hoc explanation, exacerbating tradeoffs between decisions and explanation quality: when they rely on machine predictions, their decisions are more likely to improve, but their explanations are also more likely to worsen– especially compared to relying on their own predictions and explanations.

We find evidence consistent with this interpretation. First, we observe that the negative effects of machine predictions on explanation quality is concentrated among new analysts (Figure 4(d)). Second, Figure 5(d) and Table 7 Column 3 show that treated new analysts working on simpler decisions identify more causal drivers, failing to narrow down key factors—a pattern associated with lower reasoning scores (Figure A.1(b)). They also wrote significantly shorter reports than their control counterparts (Figure A.11(d), Table A.2), suggesting more superficial reasoning.

For complex decisions, new analysts relied less on machine predictions and more on their own models (Figure A.12(a)), which may reflect the greater cognitive difficulty of constructing post hoc explanations for the machine predictions. While this reduced decision accuracy, it resulted in more coherent explanations, possibly because analysts exerted greater effort to defend their own reasoning

and argue against the algorithm, as we examine in more detail later in this section.

In contrast, experienced analysts demonstrated a stronger ability to integrate machine predictions into their reasoning. Figure 5(f)-(g)) suggest that they more effectively identified key causal drivers, and Figures A.11(f)-(g) show they produced longer, more developed explanations. This suggests they were able to focus on fewer, more important drivers and reason about them more deeply.

We find additional evidence from the explainability treatment, which presented analysts with visualized Shapley values indicating the most influential drivers behind machine predictions. New analysts working on simpler decisions were more likely to report relying on machine predictions when given Shapley explanations (Figure A.12(c)). However, they tended to incorporate machine-suggested drivers superficially, simply adding them to their list of causal factors without integrating them into a coherent story. For instance, one analyst cited a machine-suggested factor, "share of assets allocated to the top 10 holdings", as a key driver but included only a single descriptive sentence, disconnected from the rest of the explanation: "its 10 largest holdings accounted for just 24% of the portfolio as of the end of 2021, a lower clip than the Russell 1000 Index, its category benchmark." This likely contributed to the lower explanation quality we observe (Figure 4(d)).

Conversely, new analysts working on complex decisions relied even less on machine predictions when provided with explainability (Figure A.12(c)). They often challenged machine-suggested drivers and defended their own reasoning. For example, when the algorithm suggested that the "management team" was a favorable factor, a new analyst argued:

> "The management tandem is capable of executing this systematic strategy. But neither the team nor its supporting infrastructure stand out from its peers, warranting an [average] rating. ... While the management duo lacks the same pool of resources as larger, more index-oriented firms, it has capably carried out this rules-based strategy. The team is evaluated on their ability to trade efficiently and match fund performance with expectations, and they have succeeded in that respect."

In another example, a new analyst contested the machine's view that concentration was a negative factor, instead highlighting the fund's differentiated strategy:

> "The managers feel that there is considerable concentration risk in the index, as just six stocks (Apple AAPL, Amazon.com AMZN, Alphabet GOOGL, Meta FB, Tesla TSLA, and Microsoft MSFT) account for more than 40% of the Russell 1000 Growth Index benchmark as of October 2021.[The managers] are consistently underweight in that group, and the two never owned Amazon, Facebook/Meta, or Tesla. Not surprisingly, the fund's relative price metrics, such as price/earnings and price/book, are materially

> lower than those of the category average and index. ... Still, this is a sound process. While this strategy may underperform in markets where investors display an outsize appetite for lower-quality, riskier firms, its focus on long-term ownership of blue-chip companies with defensible competitive advantages remains compelling."

These examples suggest that challenging machine-suggested drivers may have enhance their reasoning by prompting analysts to argue more forcefully and provide more evidence-based arguments.

Experienced analysts, by contrast, remained selective in incorporating machine-suggested drivers and were better able to use them to strengthen their explanations. They were able to use Shapley values to pare down and identify key drivers (Figure A.9(f)). For example, an experienced analyst who received the same machine-suggested factor, "top 10 holdings", built a multi-paragraph explanation comparing the fund's concentration to benchmarks, examining its portfolio strategy, and assessing individual holdings (Appendix Section A.1.4). This deeper engagement suggests a more effective integration of machine predictions.

In sum, these findings suggest that analysts face a cognitive tradeoff when incorporating machine predictions into their decision-making. New analysts struggle to reconcile unfamiliar predictions with their own reasoning, particularly when required to construct explanations post hoc. Experienced analysts, in contrast, are more adept at critically evaluating and selectively integrating machine predictions into their mental models. This raises the possibility that experience may enable decision-makers to be better able to integrate machine predictions into their causal models, and is consistent with recent evidence that domain expertise may help decision-makers better judge predictions (Tranchero 2024).

6.1 Alternative explanations

There may also be other possible mechanisms at work. One potential explanation is that machine predictions change analysts' incentives, leading analysts to reallocate their effort to more cognitively intensive tasks (Camuffo et al. 2022). In fact, one of the company's objectives in introducing machine predictions was to enable analysts to devote more time to develop causal explanations, which they believed to be more cognitively intensive. It may also be possible that analysts were motivated to prioritize causal explanations for more complex decisions, thus reallocating their effort from simpler to more complex decisions.[17]

[17]It is worth noting that our conversations with industry experts including those within the company highlighted that explanations for simpler funds were not less important than those for complex funds. As explained by one of the experts, if analyst reasoning were less useful for simpler funds, the company would benefit from dropping analyst coverage of these funds to increase coverage of more complex funds that they lack bandwidth to cover. Experts

However, we do not observe supportive evidence. We leverage variation in the proportion of complex funds covered by analysts, as analysts' fund allocations are fixed at baseline and cannot be changed or impacted by treatment. We constructed a variable, *Share Complex*, at the analyst-month level, to measure the share of complex funds an analyst covered in each month, computed by dividing the number of complex funds covered by an analyst by the total number of funds they were assigned to cover in that month. *Share Complex* takes the value 0 when an analyst only rated simple funds in a given month, and takes the value 1 when an analyst only rated complex funds in a given month. If this effort reallocation were a key mechanism, we should see that: (1) analysts assigned to a higher proportion of complex funds deteriorate more in their reasoning on simpler funds as they have a greater opportunity to reallocate effort, and (2) analysts assigned to only simpler decisions should not deteriorate in their reasoning, as they would have no opportunity to reallocate their efforts.

We find that analysts who covered a higher proportion of complex funds were not more likely to deteriorate in the quality of their causal explanations. Results in Table A.28 Columns 1-2 focus on the subsample of analysts who were assigned only simpler funds in a month. Since these analysts do not cover any complex funds, they cannot reallocate effort from simpler funds to complex funds. However, we find that these new analysts who only cover simpler funds still exhibit significant declines in reasoning when exposed to machine predictions, which does not support the effort reallocation mechanism.

Among analysts covering a mix of simpler and complex funds in a month, we investigate whether simpler funds covered by analysts assigned to a higher share of complex funds deteriorated more in reasoning compared to those assigned to a smaller share of complex funds. In contrast to this reallocation mechanism, Table A.28 Columns 3-4 show that the causal explanations of simpler funds were not affected differentially by the share of complex funds assigned to the analysts. We also explore results in the full sample to explore whether analysts with a higher proportion of complex decisions observed their explanations deteriorate more on average. Table A.28 Columns 5-6 show that new analysts with a higher share of simpler funds show the largest declines in reasoning. This

explained that while simpler funds on average display lower volatility, the value of analyst explanations extends beyond predicting the performance of the funds. Particularly in the context of simpler funds, keeping both explicit costs (e.g., the expense ratio) and implicit costs (e.g., the cost of portfolio turnover) at a minimum is paramount (Morningstar 2018). The landscape of simpler funds is also diverse: many do not track simple and transparent indices, and make bets against the market that may risk underperformance relative to the broader market – which increases the potential role that analyst explanations about these funds can play. Consistent with these explanations, we observe that analysts do not receive higher reasoning scores on complex decisions than simpler ones on average, in the baseline data (Figure A.4).

average effect is consistent with our main results in Table 5 Columns 3-5, suggesting that new analysts suffer the most deterioration of reasoning in simpler funds.

In addition, the reallocation mechanism appears to be inconsistent with the broader results we find. While we observe that new analysts are most likely to be affected by machine predictions, Table A.27 shows that new analysts were not more likely to be assigned to simpler decisions compared to experienced analysts. This makes it unlikely that this reallocation mechanism can explain the full set of results we find.

Another potential mechanism may be that providing better machine predictions leads analysts to shirk overall. This explanation would further extend results in Dell'Acqua (2022) that when managers perceive algorithms to be of higher quality, they decrease their efforts on the prediction task by following the algorithm—raising the possibility that they may not only shirk from the prediction task but also on adjacent tasks not directly informed by the algorithm. For this mechanism to drive our results, it would imply that machine predictions were better for simpler decisions, leading analysts to shirk when working on simpler decisions – resulting in worse causal explanations.

However, we do not observe supportive evidence. To explore, we ran a comparative analysis of the algorithm's rating performance against human analyst ratings in the control group across both simple and complex funds. This comparison provides insight into the quality of machine ratings, as the control group did not observe the machine predictions and thus could not be affected by them. We compared hypothetical investment portfolios – one portfolio following analyst decisions in the control group, and the other following machine predictions. The results in Table A.30 show no statistically significant differences in the performance of the two portfolios for any decision type. This is consistent with internal evaluations within the company prior to the experiment, which showed that the accuracy of machine predictions did not vary by fund type – making it unlikely that machine predictions in the experiment were far better for simpler decisions. Moreover, given that analysts were fully aware of this information before the experiment started as they had been required to participate in multiple training sessions, it would be unlikely that they would have expected predictions to be of much higher quality for simpler decisions.

Together, these findings provide suggestive evidence that the tradeoff between decisions and causal explanations may arise because analysts face a cognitive dilemma: either construct a causal narrative around a machine prediction they do not fully understand, or rely on their own forecasts and mental models. As a result, we observe that analysts, especially those with less experience, struggle to produce coherent explanations when relying on machine predictions – highlighting that

the use of machine predictions may result in longer-term learning challenges despite short-term gains in decisions. The results further suggest that while algorithmic explainability can, in some cases, increase reliance on machine predictions, it may also amplify the tradeoffs at work.

7 Discussion and conclusion

As firms increasingly deploy algorithms to support managerial decision-making, our study highlights the importance of understanding not only how algorithms affect decisions, but also how they shape the reasoning and explanations of the decision-makers who use them. In this paper, we designed and ran a field experiment across mutual fund analysts to understand how machine predictions affect decision-makers' causal explanations in addition to the decisions they support. We find that even when algorithms improve decisions, they can deteriorate the causal explanations that decision-makers can provide for their decisions—especially those who are less experienced.

We partnered with a leading financial research firm to randomize the provision of machine predictions across analysts and decision types. This collaboration enabled us to enhance internal validity, develop deep contextual understanding, and measure difficult constructs such as the quality of causal explanations. However, the specificity of the setting may raise questions about the generalizability of our findings.

Because we study the decisions and explanations of mutual fund analysts – experts who develop deep domain knowledge through research – our findings are more likely to generalize to other expert knowledge workers. While we distinguish between relatively simpler and more complex decisions in our context, both types require significant analytical effort and domain expertise. Our results suggest that greater domain expertise may buffer experienced analysts from the trade-offs that newer analysts face when using machine predictions. Whether similar effects arise in less complex domains that do not demand specialized expertise remains an open question for future research.

Moreover, our findings are more likely to be generalizable across settings where both decisions and explanations are valuable. There are numerous contexts where decision makers are often required to not only make high quality decisions, but also offer causal explanations to support those decisions. For example, medical professionals are responsible for making recommendations and providing explanations to patients to increase the likelihood that patients can follow their recommendations (Anderson and Ledford 2024). Managers explain the reasoning behind their strategic decisions to align employees and stakeholders to execute them (Felin and Zenger 2009, 2017, Helfat

and Peteraf 2015, Mercier and Sperber 2011, Sørensen and Carroll 2021). Entrepreneurs offer investors explanations for why their business strategies will succeed (Martens et al. 2007).

Our findings are less likely to generalize to settings where causal explanations play a minimal role relative to decision accuracy. In such contexts, the tradeoff we identify between decisions and explanations – as well as the cognitive challenges of integrating machine predictions into explanations – is likely attenuated or irrelevant. For instance, air traffic controllers prioritize precision and speed over post hoc reasoning, and high-frequency trading systems operate at a scale and pace that renders individual explanations impractical and unnecessary.

More broadly, the growing use of AI raises important concerns about its unintended effects on learning and skill development (Bastani et al. 2024, Macnamara et al. 2024). Prior research has shown that algorithmic assistance can reduce effort (Dell'Acqua 2022). Our findings point to a distinct, cognitive channel: beyond affecting effort, incorporating machine-generated predictions can impair the quality of human reasoning, potentially leading to long-term skill erosion. Unlike human decision-making, where causal reasoning often precedes and evolves alongside the act of choosing, machine predictions are typically produced through opaque, correlational processes and accompanied by post hoc rationalizations. The act of crafting causal explanations – of forming hypotheses, interpreting outcomes, and revising mental models – may not be merely ancillary to decision-making but foundational to learning and long-term competence. When individuals rely on machine outputs without engaging in their own causal reasoning, they may accumulate knowledge about what works without understanding why it works (Zittrain 2019).

These concerns may become even more pronounced with the emergence of generative AI systems that can produce sophisticated natural language explanations. While our study examined relatively simple Shapley value-based feature importance scores, modern large language models can generate explanations that appear causally coherent, employ chain-of-thought reasoning, and mimic the narrative structure that professionals naturally use. This evolution in explainability methods may intensify the cognitive challenges we identify. When AI systems can produce causal narratives that emerge from fundamentally correlational learning processes, the gap between algorithmic optimization and human causal understanding becomes more obscured but not eliminated. For instance, recent work on AI-assisted creative evaluation demonstrates that while narrative AI with probabilistic rationale can improve alignment between human and algorithmic decisions, they may also lead to uncritical acceptance of AI recommendations, particularly when the AI provides plausible explanations for subjective assessments (Ayoubi et al. 2024). Such findings reinforce our concern

that decision-makers may become even more susceptible to incorporating AI-generated explanations without engaging in their own causal reasoning, potentially accelerating the erosion of the very cognitive processes that underpin expertise development. This suggests that as AI explanation capabilities advance, organizations may face increasingly subtle but profound challenges in maintaining human reasoning skills alongside improved decision accuracy.

These insights carry important implications for firms. They suggest that the benefits of AI adoption may come with hidden cognitive costs, even when we observe improvements in the short run. As organizations increasingly embed AI into their decision processes, these findings suggest that organizations should take care in how they deploy AI tools, particularly in domains where developing expertise and judgment is central to long-term performance. More broadly, they underscore the need for future research to examine how algorithmic decision aids influence the formation, refinement, and transfer of causal knowledge within firms.

References

Agarwal, R., F. Bacco, A. Camuffo, A. Coali, A. Gambardella, H. Msangi, S. T. Sonka, A. Temu, B. Waized, and A. Wormald (2023). Does a theory-of-value add value? Evidence from a randomized control trial with Tanzanian entrepreneurs. *Evidence from a Randomized Control Trial with Tanzanian Entrepreneurs (April 6, 2023)*.

Agrawal, A., J. Gans, and A. Goldfarb (2018). Prediction, judgment, and complexity: A theory of decision-making and artificial intelligence. In *The Economics of Artificial Intelligence: An Agenda*, pp. 89–110. University of Chicago Press.

Agrawal, A., J. S. Gans, and A. Goldfarb (2019, June). Exploring the impact of artificial Intelligence: Prediction versus judgment. *Information Economics and Policy 47*, 1–6.

Agrawal, A., J. McHale, and A. Oettl (2018). Finding needles in haystacks: Artificial intelligence and recombinant growth. In *The economics of artificial intelligence: An agenda*, pp. 149–174. University of Chicago Press.

Allen, R. and P. R. Choudhury (2022, January). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science 33*(1), 149–169.

Anderson, L. N. and C. J. Ledford (2024). Improving patient comprehension through explanatory communication. *JAMA 332*(23), 2027–2028.

Angrisani, M., A. Samek, and R. Serrano-Padial (2024). Competing narratives in action: An empirical analysis of model adoption dynamics. Working Paper 32242, National Bureau of Economic Research.

Armstrong, W. J., E. Genc, and M. Verbeek (2019). Going for gold: An analysis of Morningstar analyst ratings. *Management Science 65*(5), 2310–2327.

Athey, S. and G. Imbens (2017). The Econometrics of Randomized Experiments. In *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Athey, S. C., K. A. Bryan, and J. S. Gans (2020, May). The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings 110*, 80–84.

Autor, D. H. (2015, August). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives 29*(3), 3–30.

Ayoubi, C., L. Boussioux, Y. Chen, J. Ho, K. Jackson, J. Lane, C. Lin, and R. Spens (2024). The Narrative AI Advantage? A Field Experiment on Generative AI-Augmented Evaluations of Early-Stage Innovations.

Bach, L., L. E. Calvet, and P. Sodini (2020, September). Rich pickings? Risk, return, and skill in household wealth. *American Economic Review 110*(9), 2703–2747.

Barras, L., O. Scaillet, and R. Wermers (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The journal of finance 65*(1), 179–216.

Barron, K. and T. Fries (2024). Narrative persuasion.

Bastani, H., O. Bastani, A. Sungu, H. Ge, O. Kabakcı, and R. Mariman (2024). Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics. Working Paper.

Beck, M. and B. Libert (2017). The rise of AI makes emotional intelligence more important. *Harvard Business Review 15*(1-5).

Bettman, J. R. and B. A. Weitz (1983). Attributions in the board room: Causal reasoning in corporate annual reports. *Administrative Science Quarterly 28*(2), 165.

Billinger, S., N. Stieglitz, and T. R. Schumacher (2014, February). Search on rugged landscapes: An experimental study. *Organization Science 25*(1), 93–108.

Board of Governors of the Federal Reserve System (US) (2023, December). Mutual funds; Total financial assets, market value levels [BOGZ1LM654090000Q]. https://fred.stlouisfed.org/series/BOGZ1LM654090000Q.

Bollaert, H., F. Lopez-de Silanes, and A. Schwienbacher (2021). Fintech and access to finance. *Journal of corporate finance 68*, 101941.

Bradshaw, M. T., B. Lock, X. Wang, and D. Zhou (2021). Soft information in the financial press and analyst revisions. *The accounting review 96*(5), 107–132.

Brynjolfsson, E., W. Jin, and K. McElheran (2021). The power of prediction: predictive analytics, workplace complements, and business performance. *Business Economics 56*, 217–239.

Burkart, N. and M. F. Huber (2021). A Survey on the Explainability of Supervised Machine Learning. *70*, 245–317.

Campbell, D., M. Loumioti, and R. Wittenberg-Moerman (2019). Making sense of soft information: Interpretation bias and loan quality. *Journal of Accounting and Economics 68*(2-3), 101240.

Camuffo, A., A. Cordova, A. Gambardella, and C. Spina (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Science 66*(2), 564–586.

Camuffo, A., A. Gambardella, and A. Pignataro (2022). Framing strategic decisions in the digital world. *Strategic Management Review*. Forthcoming.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance 52*(1), 57–82.

Choudhury, P., E. Starr, and R. Agarwal (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal 41*(8), 1381–1411.

Coleman, M. and T. L. Liau (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology 60*(2), 283.

Cornaggia, J., K. J. Cornaggia, and H. Xia (2016). Revolving doors on wall street. *Journal of Financial Economics 120*(2), 400–419.

Cowgill, B. (2019). Bias and productivity in humans and machines. *SSRN Electronic Journal*.

Csaszar, F. A. (2018, December). What makes a decision strategic? strategic representations. *Strategy Science 3*(4), 606–619.

Dell'Acqua, F. (2022). Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Working Paper. Havard Business School, Boston.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Dietvorst, B. J., J. P. Simmons, and C. Massey (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General 144*(1), 114–126.

Ehrig, T. and J. Schmidt (2021). Making biased but better predictions: The trade-offs strategists face when they learn and use heuristics. *Strategic Organization 19*(2), 263–284.

Eliaz, K. and R. Spiegler (2020). A model of competing narratives. *The American Economic Review 110*(12), pp. 3786–3816.

Ertugrul, M. and S. Hegde (2009, March). Corporate governance ratings and firm performance. *Financial Management 38*(1), 139–160.

Ethiraj, S. K. and D. Levinthal (2004, February). Modularity and innovation in complex systems. *Management Science 50*(2), 159–173.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of financial economics 116*(1), 1–22.

Felin, T. and T. R. Zenger (2009). Entrepreneurs as theorists: on the origins of collective beliefs and novel strategies. *Strategic Entrepreneurship Journal 3*(2), 127–146.

Felin, T. and T. R. Zenger (2017). The theory-based view: economic actors as theorists. *Strategy Science 2*(4), 258–271.

Fernandez, R. and D. Rodrik (1991). Resistance to reform: Status quo bias in the presence of individual-specific uncertainty. *The American economic review*, 1146–1155.

Fiss, P. C. and E. J. Zajac (2006). The symbolic management of strategic change: Sensegiving via framing and decoupling. *Academy of management journal 49*(6), 1173–1193.

Goetzmann, W. N. and N. Peles (1997). Cognitive dissonance and mutual fund investors. *Journal of financial Research 20*(2), 145–158.

Grennan, J. and R. Michaely (2020). Artificial intelligence and high-skilled work: Evidence from analysts. *Swiss Finance Institute Research Paper* (20-84).

Gruber, M. J. (1996). Another puzzle: The growth in actively managed mutual funds. *The journal of finance 51*(3), 783–810.

Helfat, C. E. and M. A. Peteraf (2015, June). Managerial cognitive capabilities and the microfoundations of dynamic capabilities. *Strategic Management Journal 36*(6), 831–850.

Hoffman, M., L. B. Kahn, and D. Li (2018, May). Discretion in hiring. *The Quarterly Journal of Economics 133*(2), 765–800.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945-1964. *The Journal of finance 23*(2), 389–416.

Kempf, E. (2020). The job rating game: Revolving doors and analyst incentives. *Journal of Financial Economics 135*(1), 41–67.

Kendall, C. W. and C. Charles (2022). Causal narratives. Working Paper 30346, National Bureau of Economic Research.

Kim, H., E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca (2024). Decision authority and the returns to algorithms. *Strategic Management Journal 45*(4), 619–648.

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training, University of Central Florida.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017, August). Human decisions and machine predictions. *The Quarterly Journal of Economics*.

Lebovitz, S., H. Lifshitz-Assaf, and N. Levina (2022, January). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science 33*(1), 126–148.

Lou, B. and L. Wu (2021, September). AI on drugs: Can artificial intelligence accelerate drug development? Evidence from a large-scale examination of bio-pharma firms. *MIS Quarterly 45*(3), 1451–1482.

Lourie, B. (2019). The revolving door of sell-side analysts. *The Accounting Review 94*(1), 249–270.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems 30*.

Luo, X., M. S. Qin, Z. Fang, and Z. Qu (2021). Artificial intelligence coaches for sales agents: Caveats and solutions. *Journal of Marketing 85*(2), 14–32.

Luo, X., H. Wang, S. Raithel, and Q. Zheng (2015, January). Corporate social performance, analyst stock recommendations, and firm future returns: Research Notes and Commentaries. *Strategic Management Journal 36*(1), 123–136.

MacLennan, A. F. and C. C. Markides (2021). Causal mapping for strategy execution: Pitfalls and applications. *California Management Review 63*(4), 89–122.

Macnamara, B. N., I. Berber, M. C. Çavuşoğlu, E. A. Krupinski, N. Nallapareddy, N. E. Nelson, P. J. Smith, A. L. Wilson-Delfosse, and S. Ray (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications 9*(1), 46.

Martens, M. L., J. E. Jennings, and P. D. Jennings (2007). Do the Stories They Tell Get Them the Money They Need? The Role of Entrepreneurial Narratives in Resource Acquisition. *The Academy of Management Journal 50*(5), 1107–1132.

Mercier, H. and D. Sperber (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences 34*(2), 57–74.

Morningstar (2018). Morningstar's guide to passive investing. Report, Morningstar.

Pfeffer, J. (1981). Management as symbolic action: the creation and maintenance of organizational paradigm. *Research in organizational behavior 3*, 1–52.

Raisch, S. and S. Krakowski (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of management review 46*(1), 192–210.

Rudin, C. (2019, May). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*(5), 206–215.

Röder, M., A. Both, and A. Hinneburg (2015). Exploring the space of topic coherence measures. pp. 399–408.

Samuelson, W. and R. Zeckhauser (1988). Status quo bias in decision making. *Journal of risk and uncertainty 1*, 7–59.

Schmitt, W. (2024, 18 Jan). Passive eclipses active in us fund market as assets swell to $13.3tn. *Financial Times*. Available at: `https://www.ft.com/content/faf74f66-c4d6-45aa-bb30-0e73d523c547` (Accessed: March 1st, 2024).

Schoar, A. and Y. Sun (2024). Financial advice and investor beliefs: Experimental evidence on active vs. passive strategies. Working Paper 33001, National Bureau of Economic Research.

Shaw, W. and A. Gani (2023, May). Wall street banks are using AI to rewire the world of finance. *Bloomberg.com*.

Shrestha, Y. R., S. M. Ben-Menahem, and G. Von Krogh (2019). Organizational decision-making structures in the age of artificial intelligence. *California management review 61*(4), 66–83.

Siggelkow, N. (2002, July). Misperceiving interactions among complements and substitutes: Organizational consequences. *Management Science 48*(7), 900–916.

Simon, H. (1947). *Administrative behavior: A study of decision-making processes in administrative organizations*. New York: Macmillan.

Sirri, E. R. and P. Tufano (1998). Costly search and mutual fund flows. *The journal of finance 53*(5), 1589–1622.

Sørensen, J. B. and G. R. Carroll (2021, June). Why Good Arguments Make Better Strategy. *MIT Sloan Management Review 62*(4).

Soydemir, G., J. Smolarski, and S. Shin (2014). Hedge funds, fund attributes and risk adjusted returns. *Journal of Economics and Finance 38*, 133–149.

Stewart, A. (2020, June). The former amazon vp who quit in solidarity with fired employees published a business case for why he thinks the massively profitable amazon web services cloud business should be spun off. Accessed: September 30, 2024.

Tomaino, G., H. Abdulhalim, P. Kireyev, and K. Wertenbroch (2020). Denied by an (Unexplainable) Algorithm: Teleological Explanations for Algorithmic Decisions Enhance Customer Satisfaction. *SSRN Electronic Journal*.

Tong, S., N. Jia, X. Luo, and Z. Fang (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal 42*(9), 1600–1631.

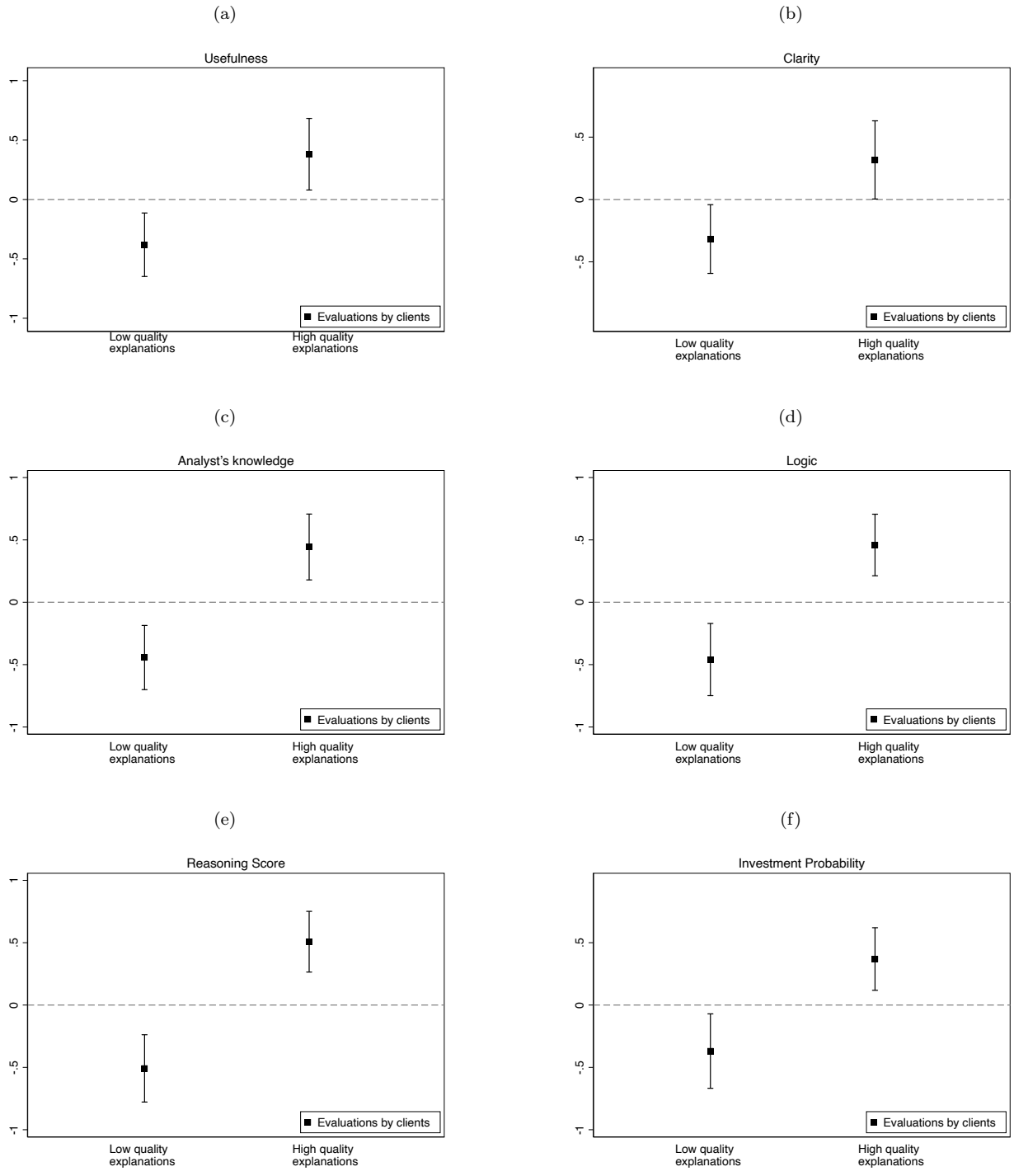
Tranchero, M. (2024, August). Finding Diamonds in the Rough: Data-Driven Opportunities and Pharmaceutical Innovation. *Academy of Management Proceedings 2024*(1), 13751.

Vaccaro, M., A. Almaatouq, and T. Malone (2019). When combinations of humans and AI are useful: A systematic review and meta-analysis. *8*(12), 2293–2303.

Zittrain, J. (2019). The Hidden Costs of Automated Thinking. *The New Yorker*.

Figures and Tables

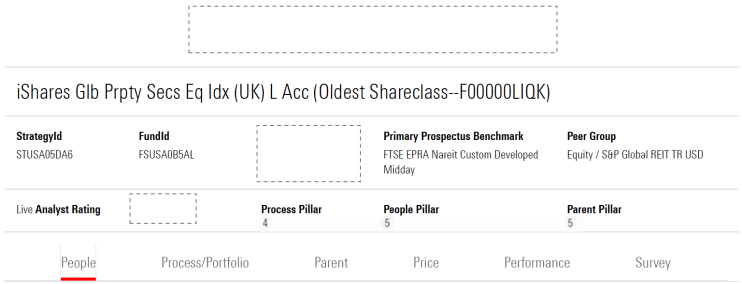
Figure 1: Validating reasoning scores with external financial advisors



Notes: These figures show the mean and confidence interval plots of external financial advisors' evaluation of analysts' explanations along the dimensions of usefulness, clarity, knowledge, logic, reasoning score, and the probability they will invest in the recommended fund after reviewing the analysts' explanations, split by whether the explanations were previously rated as high quality by internal review committees, our primary measure of explanation quality. "Low quality explanations" indicate reports that received below-median reasoning scores from the internal committee; "high quality explanations" indicate those that received above-median reasoning scores. All funds in the sample of this survey experiment received the same analyst rating, which means that all funds in the study were recommended by analysts to the same degree. The confidence intervals are calculated at 95% level.

Figure 2: Design of control and treatment conditions

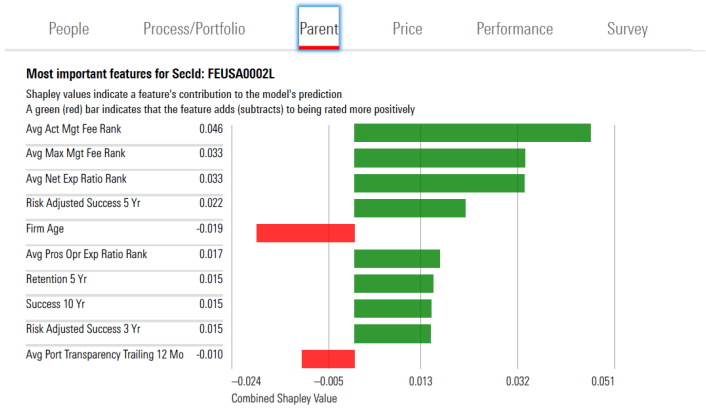
(a) Control



(b.1) Treatment

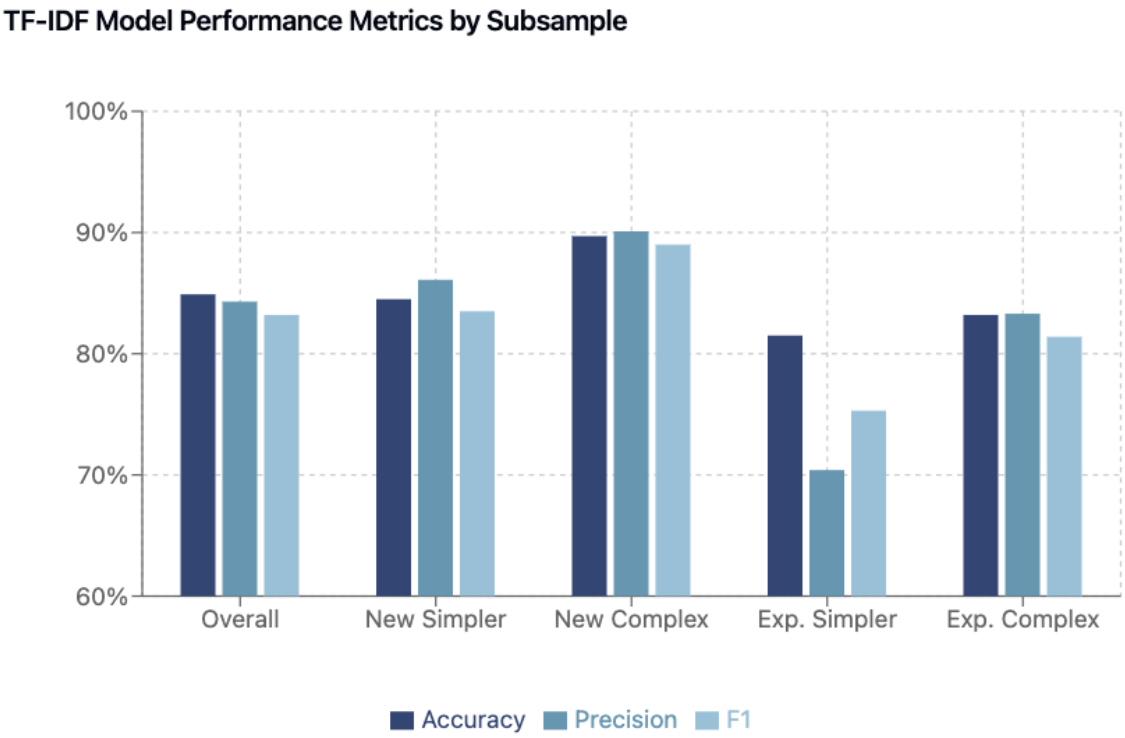
(b.2) Treatment cont.

(c) Algorithm explainability using Shapley values



Notes: This set of figures shows what analysts saw on the research platform across different experimental conditions. The control group continued to do research as usual, and saw no changes on the interface ((a)). Analysts assigned to either treatment group saw an additional line of information at the top of the interface that provided a machine prediction for the fund ((b.1-2)). (a) shows an example of a non-actively managed fund, while (b) shows an example of an active equity fund. Analysts in treatment group 1 received algorithmic predictions only, while treatment 2 additionally saw a figure that highlighted the ten key features that contributed most to the algorithm's prediction for that fund ((c)), which was based on Shapley values.

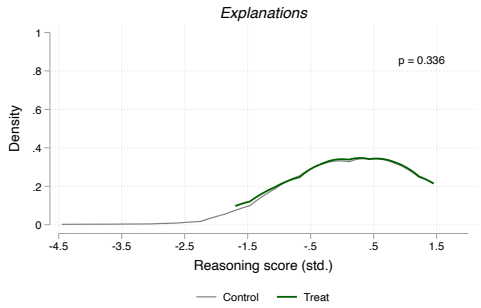
Figure 3: Performance of TF-IDF model that classifies analyst reports into treatment conditions



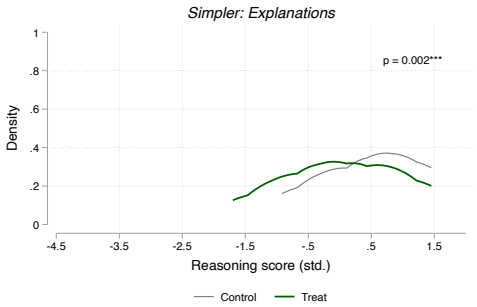
Notes: This figures present the performance of textual analysis model by subsample. We train a Light Gradient-Boosting Machine (LightGBM) model with TF-IDF vectorization to evaluate whether the treatment conditions an analyst was assigned to could be predicted from the analysts' text alone.

Figure 4: The effects of machine predictions on explanations by analyst experience and decision type

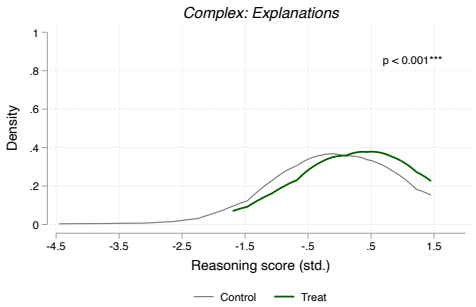
(a)



(b)

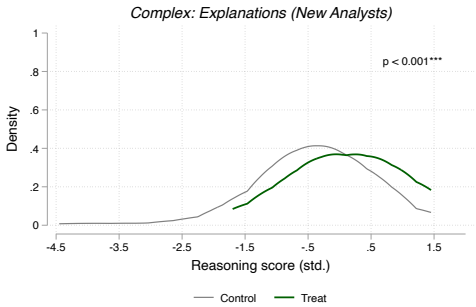


(c)

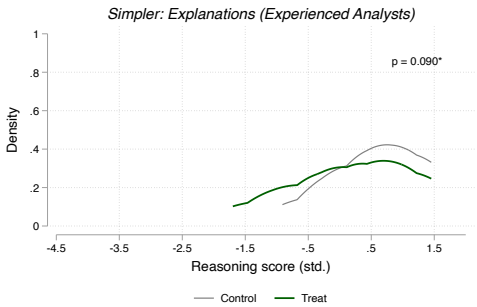


(d)

(e)



(f)

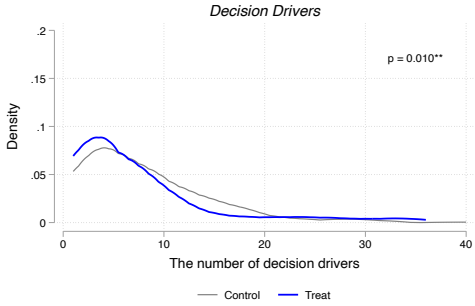


(g)

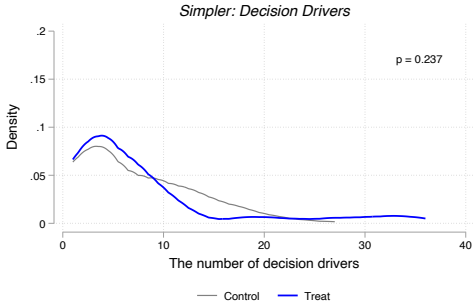
Notes: These figures show kernel density plots of the quality of explanations by treatment and control conditions on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* are included in the plots. * p<0.1, ** p<0.05, *** p<0.01.

Figure 5: The effects of machine predictions on the number of decision drivers

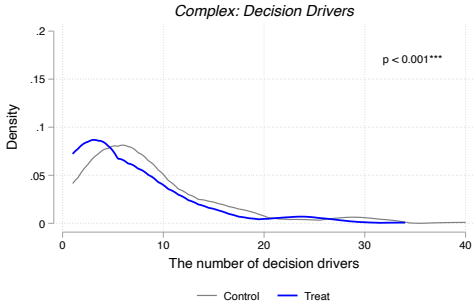
(a)



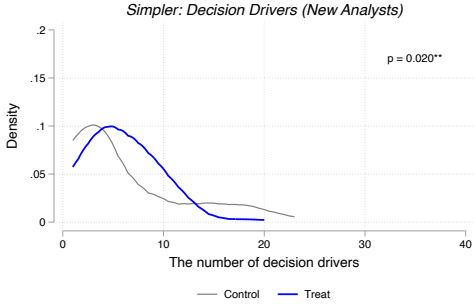
(b)



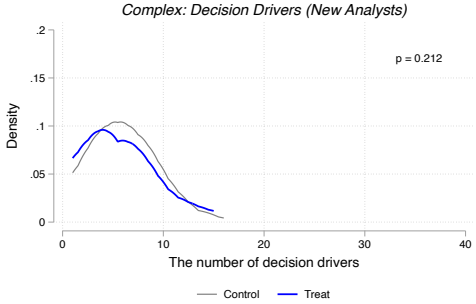
(c)



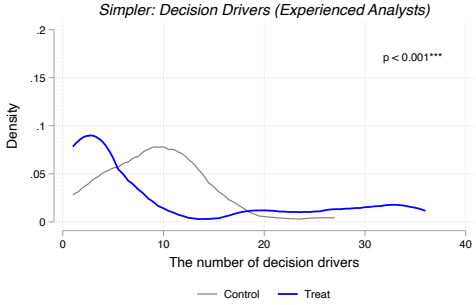
(d)



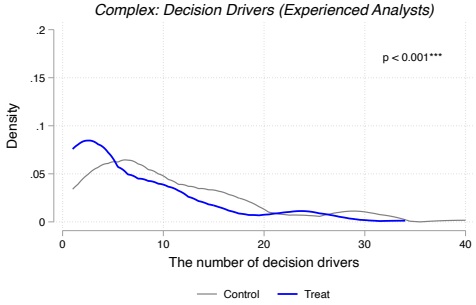
(e)



(f)



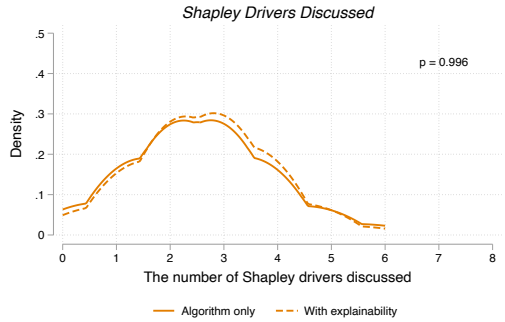
(g)



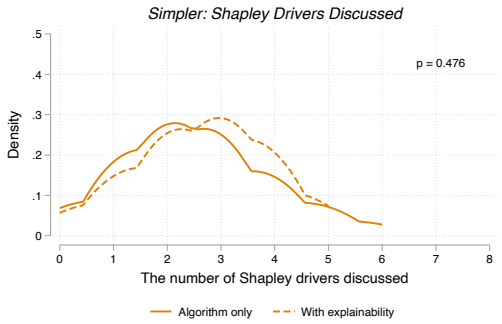
Notes: These figures show kernel density plots of the number of decision drivers by treatment and control conditions on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* are included in the plots. * p<0.1, ** p<0.05, *** p<0.01.

Figure 6: The effects of machine explainability on the number of machine-suggested Shapley drivers discussed by decision type and analyst experience

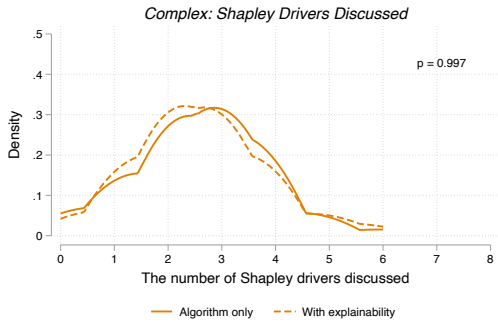
(a)



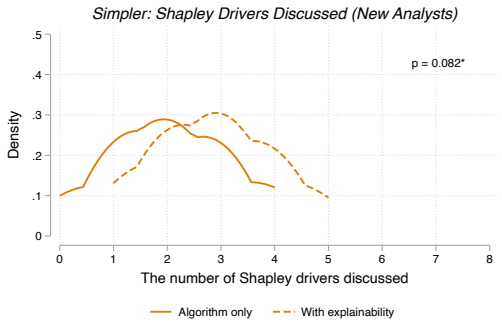
(b)



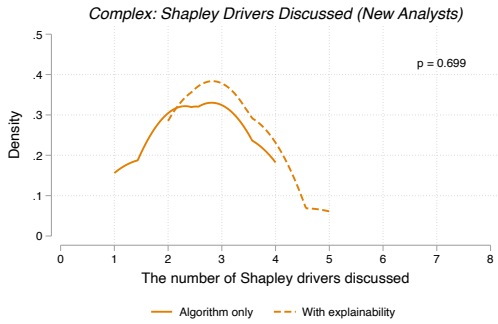
(c)



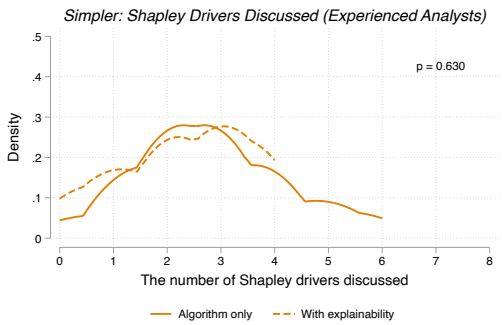
(d)



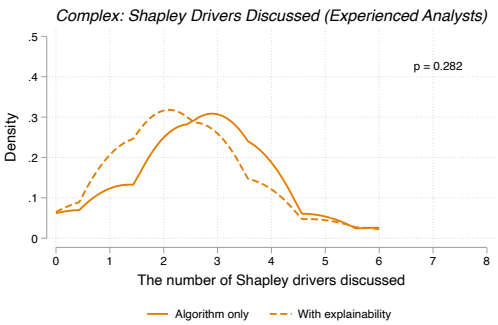
(e)



(f)



(g)



Notes: These figures show kernel density plots of the number of machine-suggested Shapley drivers discussed in analyst reports by treatment condition with machine prediction only and treatment condition with both machine predictions and explainability on average ((a)) and by subgroups ((b)-(g)).

Table 1: Validating reasoning scores with external financial advisors

	Usefulness	Clarity	Knowledge	Logic	Reasoning Score	Investment Probability
	(1)	(2)	(3)	(4)	(5)	(6)
Reasoning score (std.)	0.251***	0.222**	0.310***	0.350***	0.376***	0.289***
	(0.085)	(0.086)	(0.079)	(0.079)	(0.077)	(0.080)
Observations	64	64	64	64	64	64
Mean	0.000	0.000	0.000	-0.000	-0.000	0.000
SD	0.873	0.873	0.845	0.873	0.873	0.845

Notes: This table presents the robustness check for reasoning scores rated by internal review committees. In a follow-up experiment, 16 external financial advisors evaluated a randomly drawn set of 4 analyst reports from a pool of 20 on analysts' explanations, which allows us to compare their evaluation of analyst explanations with reasoning scores rated by internal review committees. Column (1)-(6) show the OLS estimates of the correlation between the internal experts' standardized reasoning scores and the external experts' standardized rating of the analyst explanations along the dimensions of usefulness, clarity, knowledge, logic, reasoning (i.e., the combination of knowledge and logic), and the probability they will invest in a fund after reading the analysts' recommendation and explanations. External expert financial advisors rated each dimension on a scale of 1 to 10, except for the investment probability, which was rated on a scale of 1 to 100. The ratings are standardized at the financial advisor level. Standard errors are in parentheses, clustered at the analyst level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 2: Summary statistics and balance table

	Full Sample				Control	Treatment	Difference	
	Mean	SD	Min.	Max.	Mean	Mean	p-value	RI p-value
Panel A: Analyst's fund coverage								
Percentage of complex funds	0.483	0.468	0.000	1.000	0.498	0.466	0.372	0.743
Percentage of recommendations	0.550	0.301	0.000	1.000	0.540	0.561	0.630	0.737
Average fund age	19.743	9.254	6.622	71.540	20.642	18.762	0.166	0.350
Average fund size	2.627	3.741	0.000	20.879	2.732	2.509	0.391	0.781
Number of funds covered	20.043	10.636	1.000	66.000	19.208	20.955	0.783	0.435
Percentage of rating change	0.112	0.133	0.000	0.429	0.133	0.089	0.062	0.115
Reasoning score (std.)	-0.085	0.639	-1.995	1.124	-0.047	-0.127	0.278	0.557
Fund returns in 3 months	-4.131	8.700	-33.434	31.432	-5.315	-2.840	0.913	0.171
Fund returns in 6 months	-4.947	6.296	-19.003	14.749	-5.143	-4.732	0.622	0.751
Number of decision drivers	6.281	4.903	1.000	21.316	6.918	5.661	0.129	0.264
Number of causal statements	8.206	1.408	5.000	13.000	8.356	8.036	0.217	0.442
Panel B: Analyst characteristics								
Experienced	0.649	0.480	0.000	1.000	0.673	0.625	0.311	0.691
Tenure	7.008	6.186	0.647	29.490	6.859	7.160	0.594	0.811
Time in current position	6.224	6.021	0.647	29.490	5.984	6.468	0.653	0.695
Male	0.701	0.460	0.000	1.000	0.673	0.729	0.723	0.656
Manager	0.093	0.292	0.000	1.000	0.102	0.083	0.377	1.000
Professional certificate	0.309	0.465	0.000	1.000	0.347	0.271	0.211	0.510
Postgraduate degree	0.433	0.498	0.000	1.000	0.429	0.438	0.535	1.000
Observations	97				49	48	97	97

Notes: This table reports summary statistics of the full sample and the OLS estimates of baseline differences between control and treatment groups. For each variable, the table reports the full sample mean, SD, minimum and maximum. It also reports the baseline control group mean and treatment group mean. The table reports the p-value for the t-test between treatment group and control group, as well as the randomization inference (RI) p-value for the treatment indicator based on 5000 draws.
Percentage of complex funds measures the proportion of complex funds relative to the total number of funds that an analyst is assigned to cover during the observation window. Percentage of recommendations measures the proportion of decisions to recommend a fund relative to the total number of decisions made by an analyst during the observation window. *Fund age* is calculated as the number of years since the fund's inception. *Fund size* is approximated by total net assets under management (in billions of USD). *Number of funds covered* counts the total number of funds assigned to an analyst during the observation period. *Percentage of rating change* measures the proportion of decision to change the live rating, as oppose to maintaining the status quo, relative to the total number of decisions made by an analyst. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. *Fund returns in 3 months* measures the risk-adjusted return of a fund 3 months after its rating is published. *Fund returns in 6 months* measures the risk-adjusted return of a fund 6 months after its rating is published. *Decision drivers* measure the number of reported decision drivers by the analyst and is natural log transformed. *Causal statements* measure the number of causal statements in the analyst report as identified by GPT-4 and is natural log transformed. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Tenure* measures the tenure of analysts in years. *Time in current position* counts the number of years since an analyst started working in their current position. *Male* is an indicator variable that takes the value 1 if the analyst is male and 0 otherwise. *Manager* is an indicator variable that takes the value 1 if the analyst holds managerial position and 0 otherwise. *Professional certificate* takes the value 1 if the analyst holds at least one professional certificates, such as CPA and CFA, and 0 otherwise. *Postgraduate degree* takes the value 1 if the analyst holds postgraduate degree and 0 otherwise.

Table 3: Details on missing data

	Treatment		Control		p-value
	Number of observations	% of missing observations	Number of observations	% of missing observations	
Reasoning score	608	32.294	664	24.717	0.258
Fund returns in 3 months	892	0.668	873	1.02	0.842
Fund returns in 6 months	892	0.668	870	1.361	0.687

Notes: This table reports the summary statistics on missing data on the main outcome variables, *Reasoning score* and *fund returns* measures, by treatment and control conditions. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. The *fund returns in 3 months* measures the risk-adjusted return of a fund 3 months after its rating is published. The *fund returns in 6 months* measures the risk-adjusted return of a fund 6 months after its rating is published. The table reports the p-value for the t-test between treatment group and control group.

Table 4: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating					
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.074**	0.075**	0.080**	0.080**	0.080*	0.114**
	(0.033)	(0.033)	(0.031)	(0.031)	(0.043)	(0.057)
	[0.029]	[0.030]	[0.016]	[0.017]	[0.084]	[0.057]
Treat × Post × Complex					-0.000	-0.039
					(0.071)	(0.122)
					[1.000]	[0.773]
Treat × Post × Experienced						-0.042
						(0.084)
						[0.636]
Treat × Post × Complex × Experienced						0.043
						(0.151)
						[0.785]
Strata FE	No	Yes	No	Yes	Yes	Yes
Month FE	No	No	Yes	Yes	Yes	Yes
Observations	1780	1780	1780	1780	1780	1780
Mean (control)	0.113	0.113	0.113	0.113	0.113	0.113
SD (control)	0.317	0.317	0.317	0.317	0.317	0.317

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.894	9.886***	11.557***	1.130	7.949***	6.741*
	(3.033)	(2.585)	(3.600)	(2.165)	(2.487)	(3.761)
	[0.792]	[0.001]	[0.010]	[0.625]	[0.004]	[0.154]
Treat × Post × Recommend × Complex		-16.822***	-16.747**		-11.458***	-8.323
		(5.268)	(7.857)		(4.191)	(6.471)
		[0.003]	[0.063]		[0.012]	[0.228]
Treat × Post × Recommend × Experienced			-5.959			-0.604
			(5.031)			(4.510)
			[0.328]			[0.923]
Treat × Post × Recommend × Complex			5.281			-0.092
× Experienced			(10.532)			(8.418)
			[0.660]			[0.990]
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.188	-4.188	-4.188	-4.243	-4.243	-4.243
SD (control)	13.620	13.620	13.620	12.812	12.812	12.812

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include a full set of strata and month fixed effects. Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at the analyst level. Random inference p-values are in square brackets, calculated using 5000 repetitions. Standard errors are clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table 5: The impact of machine predictions on explanations

	All committees				Single
	(1)	(2)	(3)	(4)	(5)
Treat × Post	0.012	-0.124	-0.661*	-0.564*	-0.698**
	(0.169)	(0.210)	(0.345)	(0.322)	(0.339)
	[0.945]	[0.582]	[0.060]	[0.053]	[0.029]
Treat × Post × Complex		0.297	0.907**	0.609	1.151**
		(0.351)	(0.395)	(0.387)	(0.446)
		[0.407]	[0.031]	[0.178]	[0.020]
Treat × Post × Experienced			0.828*	0.668*	0.786*
			(0.438)	(0.374)	(0.398)
			[0.063]	[0.063]	[0.056]
Treat × Post × Complex × Experienced			-0.950	-0.796*	-1.293**
			(0.607)	(0.465)	(0.529)
			[0.139]	[0.141]	[0.038]
Strata FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes
Observations	1313	1313	1313	1207	901
Mean (control)	0.238	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.928	0.968

Notes: All regressions include a full set of strata and month fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include reasoning scores issued by committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of reasoning scores issued by committees with one reviewer only. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. Random inference p-values are in square brackets, calculated using 5000 repetitions. Standard errors are clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table 6: The impact of machine explainability on decisions and explanations

Panel A. Decision change

	Change rating			Reasoning score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.120***	0.097**	0.081	0.005	-0.362	-0.788**
	(0.041)	(0.048)	(0.063)	(0.235)	(0.223)	(0.358)
Explainability × Post	0.040	0.062	0.151**	0.020	0.112	-0.554
	(0.034)	(0.050)	(0.064)	(0.194)	(0.278)	(0.407)
Treat × Post × Complex		0.057	0.005		0.736	0.981**
		(0.098)	(0.152)		(0.453)	(0.434)
Explainability × Post × Complex		-0.049	-0.025		-0.181	0.984**
		(0.069)	(0.093)		(0.379)	(0.391)
Treat × Post × Experienced			0.063			0.496
			(0.093)			(0.424)
Treat × Post × Complex × Experienced			0.051			-0.246
			(0.186)			(0.804)
Explainability × Post × Experienced			-0.150			1.099*
			(0.094)			(0.556)
Explainability × Post × Complex ×			0.017			-1.666***
Experienced			(0.128)			(0.626)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1780	1780	1780	1313	1313	1313
Mean (control)	0.113	0.113	0.113	0.238	0.238	0.238
SD (control)	0.317	0.317	0.317	0.928	0.928	0.928

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.002	10.461***	15.499***	1.834	8.679***	9.799**
	(4.065)	(3.656)	(4.443)	(2.547)	(3.195)	(4.478)
Explainability × Post × Recommend	2.235	10.209***	11.637**	0.322	7.318**	3.152
	(3.357)	(3.347)	(5.127)	(3.118)	(3.272)	(4.878)
Treat × Post × Recommend × Complex		-18.858***	-20.091**		-10.709**	-12.275
		(6.436)	(9.755)		(5.026)	(7.462)
Explainability × Post × Recommend ×		-15.289**	-19.980***		-12.109**	-2.050
Complex		(6.671)	(6.769)		(6.028)	(7.529)
Treat × Post × Recommend × Experienced			-10.428*			-3.031
			(6.013)			(5.443)
Treat × Post × Recommend × Complex			5.981			6.060
× Experienced			(11.883)			(9.287)
Explainability × Post × Recommend ×			-7.221			0.178
Experienced			(8.201)			(6.744)
Explainability × Post × Recommend ×			12.747			-5.065
Complex × Experienced			(11.874)			(11.021)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.188	-4.188	-4.188	-4.243	-4.243	-4.243
SD (control)	13.620	13.620	13.620	12.812	12.812	12.812

Notes: Panel A shows treatment effects on *Change rating* and on *Reasoning score*. Panel B shows whether recommended funds by analysts in different treatment conditions observe higher fund returns in subsequent months. *Treat* is an indicator variable that takes the value 1 for analysts assigned to treatment 1 who receive only the algorithmic prediction. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 who receive algorithmic predictions along with key fund features that contributed most to the prediction based on Shapley values. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table 7: The number of decision drivers and causal statements by experimental condition

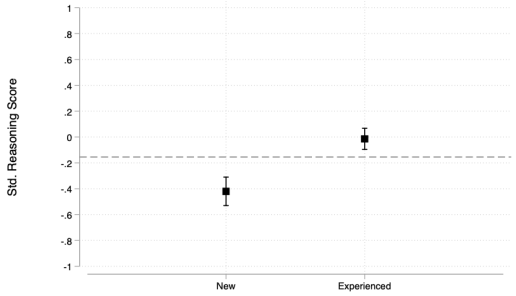
	Decision Drivers			Causal Statements		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	-0.332*	0.033	0.478	-0.065	-0.114**	0.026
	(0.187)	(0.259)	(0.321)	(0.040)	(0.052)	(0.082)
	[0.114]	[0.906]	[0.267]	[0.136]	[0.062]	[0.743]
Treat × Post × Complex		-0.610*	-1.378**		0.110	-0.194*
		(0.352)	(0.542)		(0.080)	(0.111)
		[0.118]	[0.056]		[0.245]	[0.221]
Treat × Post × Experienced			-0.893*			-0.240**
			(0.459)			(0.104)
			[0.154]			[0.068]
Treat × Post × Complex × Experienced			1.331**			0.496***
			(0.669)			(0.142)
			[0.146]			[0.013]
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1317	1317	1317	948	948	948
Mean (control)	1.728	1.728	1.728	2.104	2.104	2.104
SD (control)	0.896	0.896	0.896	0.274	0.274	0.274

Notes: All regressions include a full set of strata and month fixed effects. *Decision drivers* measure the number of reported decision drivers by the analyst and is natural log transformed. *Causal statements* measure the number of causal statements in the analyst report as identified by GPT-4 and is natural log transformed. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. Random inference p-values are in square brackets, calculated using 5000 repetitions. Standard errors are clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

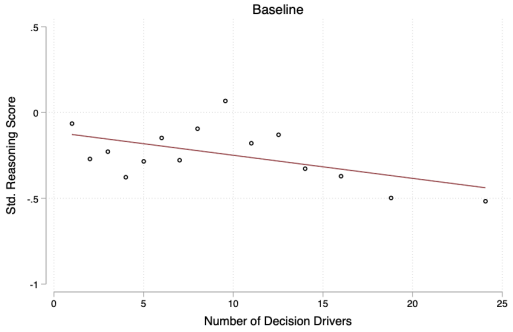
Appendix

Figure A.1: Baseline descriptives on causal explanations

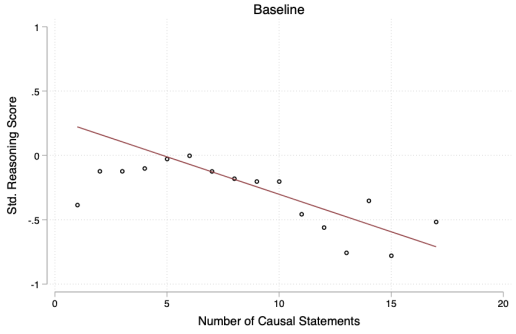
(a) Baseline reasoning scores by analyst experience



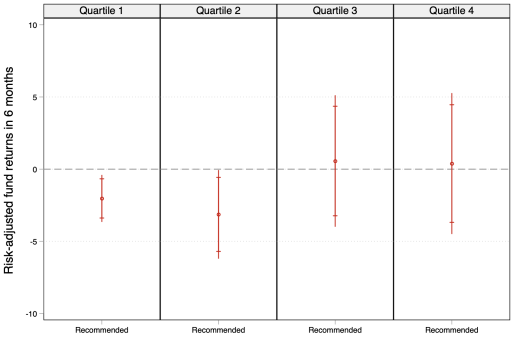
(b) Baseline correlation between reasoning scores and decision drivers



(c) Baseline correlation between reasoning scores and causal statements

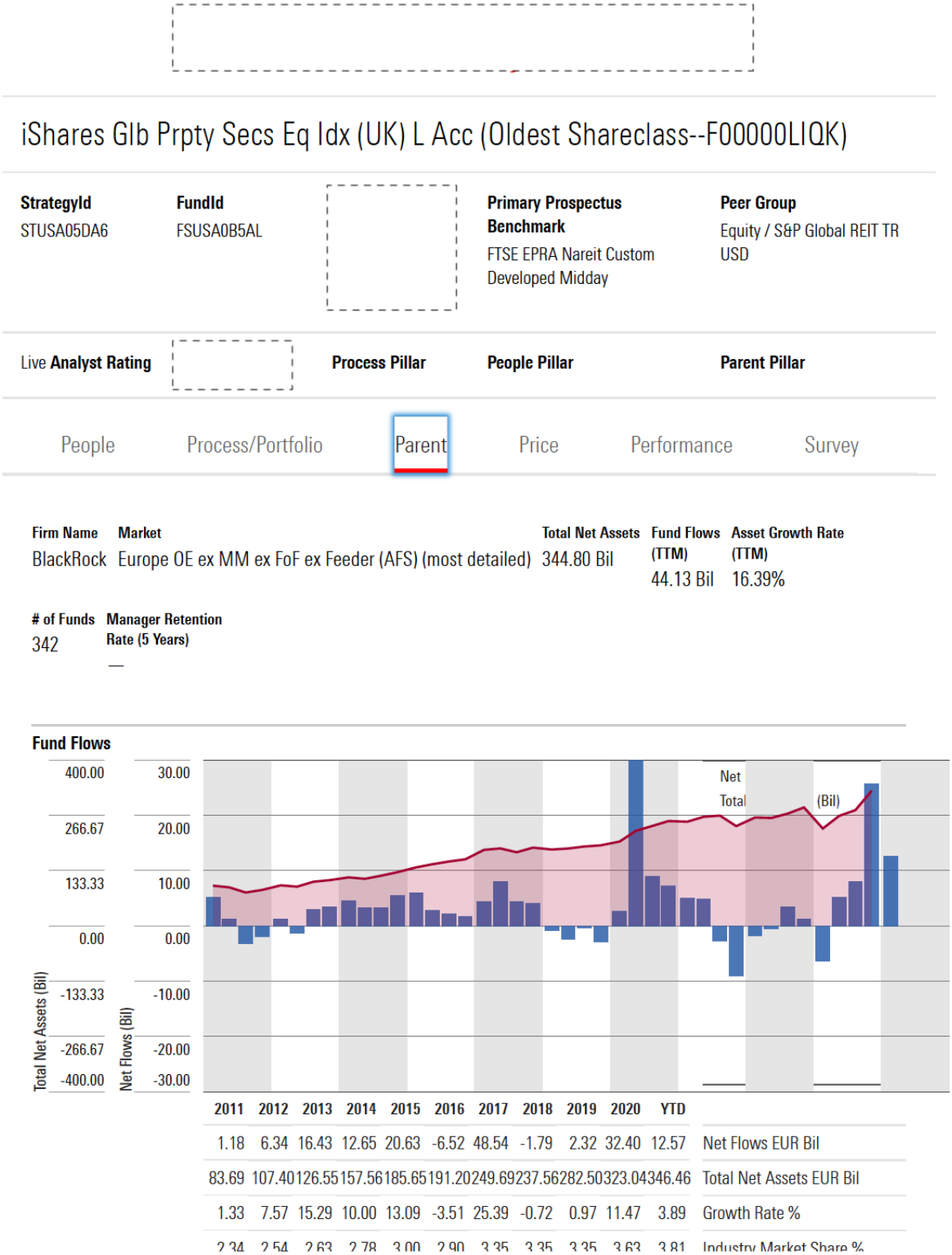


(d) Fund returns by quartiles of reasoning scores



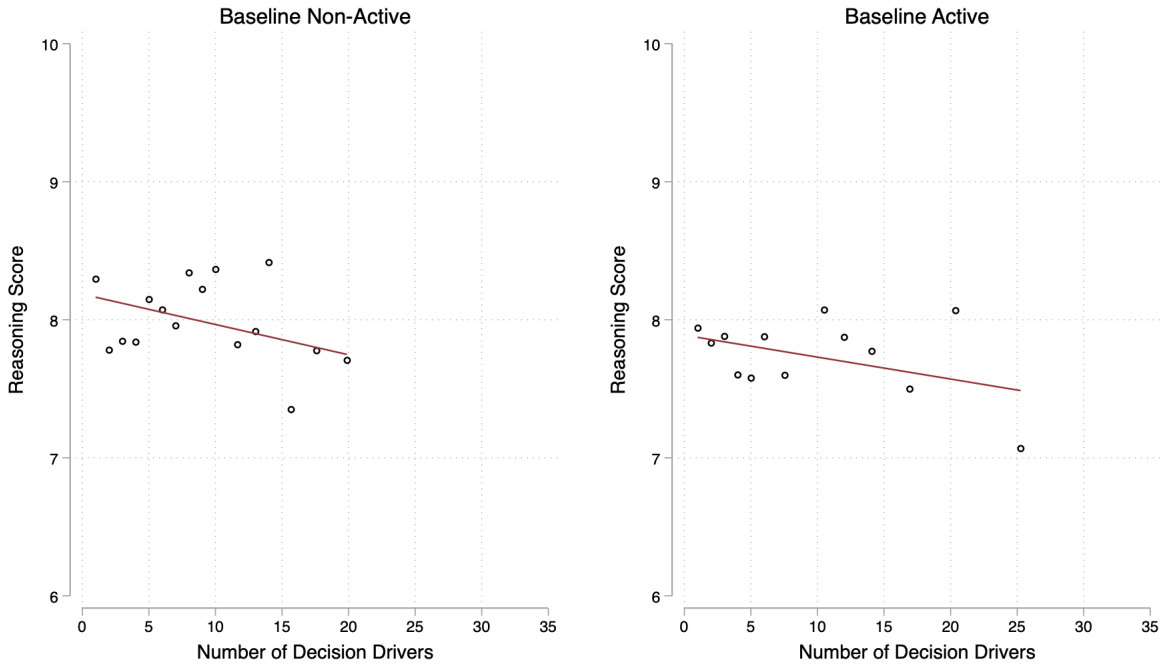
Notes: Figure (a) shows the relationship between reasoning scores and the number of decision drivers at baseline. *Reasoning score* is constructed by standardizing reasoning scores (evaluated on a scale of 1 to 10) across individual expert committee members and taking an average across them for each fund. The *number of decision drivers* counts the number of factors that analysts report as driving their rating decisions. Figure (b) shows the average reasoning score by analyst experience. Experienced analysts are those that have more than three years of rating experience at the company.

Figure A.2: Online interface for analysts



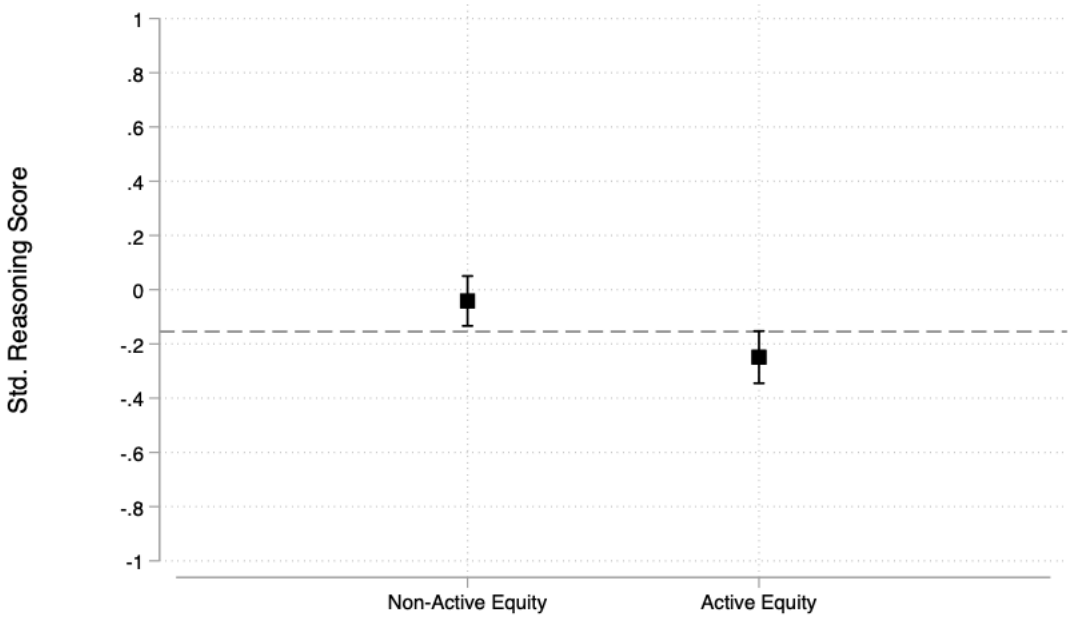
Notes: This figure shows a screenshot of the online research interface where analysts performed their work. Analysts saw the fund's current live rating from the last evaluation and received information about the fund for their analysis. On this interface, analysts decided on an up-to-date rating for the fund and delineated their causal explanation for the rating.

Figure A.3: Baseline correlation between reasoning scores and decision drivers by fund type



Notes: The figures show binned scatterplots of reasoning scores on the number of decision drivers at baseline by fund type (simpler NAE funds or complex AE funds). *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. The *number of decision drivers* counts the number of factors that analysts report as driving their rating decisions.

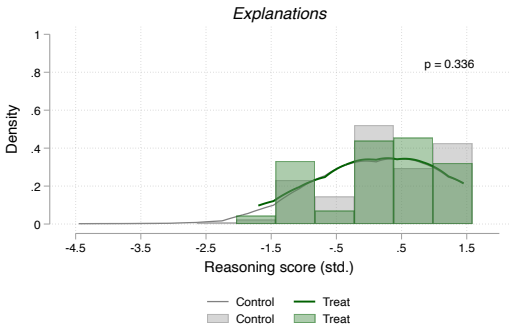
Figure A.4: Baseline reasoning scores by fund type



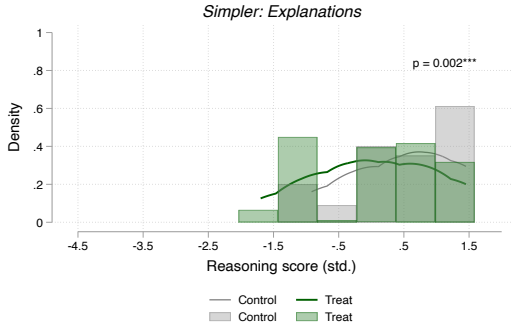
Notes: This figure shows means and 95% confidence intervals of reasoning scores by fund type at baseline. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Fund type is separated into simpler funds and complex funds.

Figure A.5: The effects of machine predictions on explanations by decision type and analyst experience

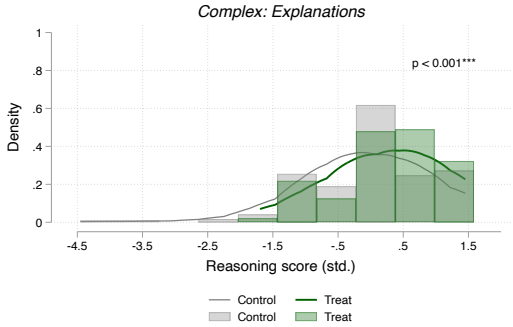
(a)



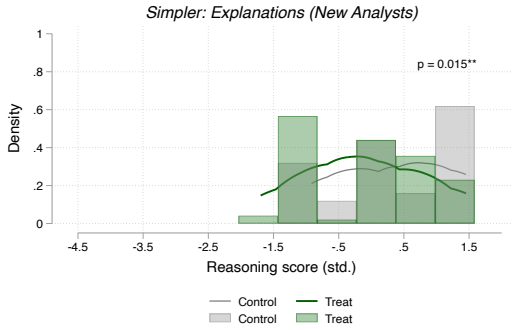
(b)



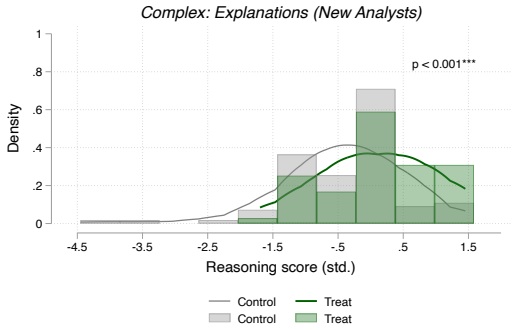
(c)



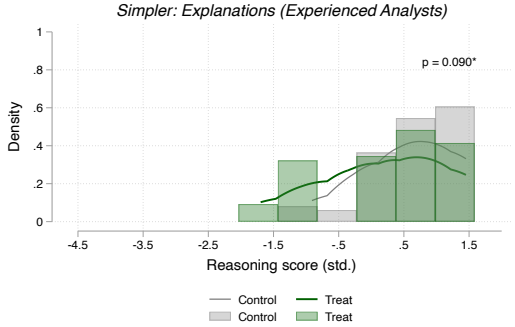
(d)



(e)



(f)

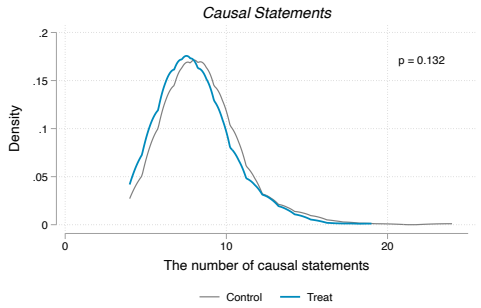


(g)

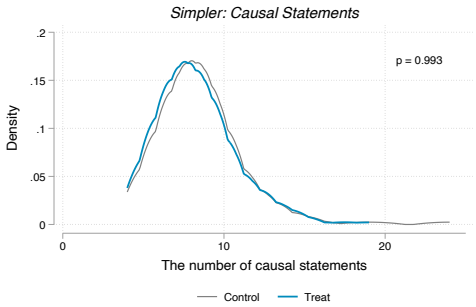
Notes: These figures show kernel density plots and histograms of reasoning scores by treatment and control conditions on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* are included in the plots. * p<0.1, ** p<0.05, *** p<0.01.

Figure A.6: The effects of machine predictions on the number of causal statements by decision type and analyst experience

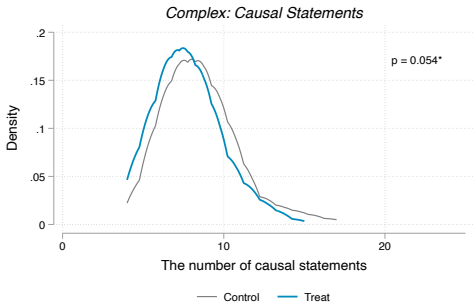
(a)



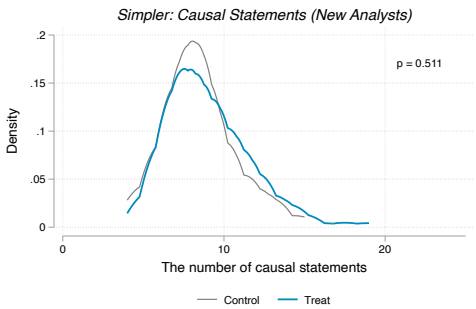
(b)



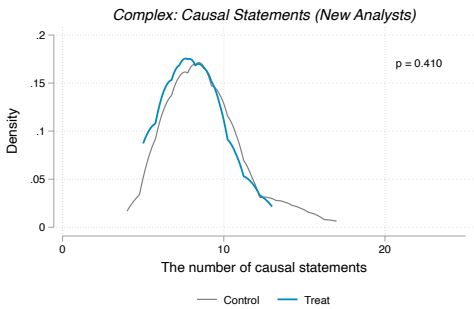
(c)



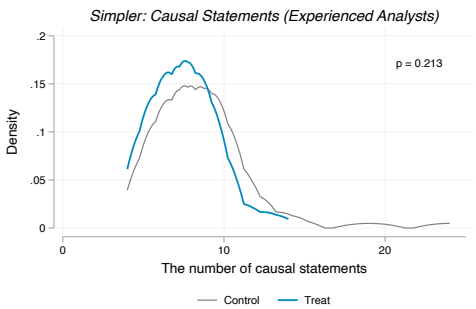
(d)



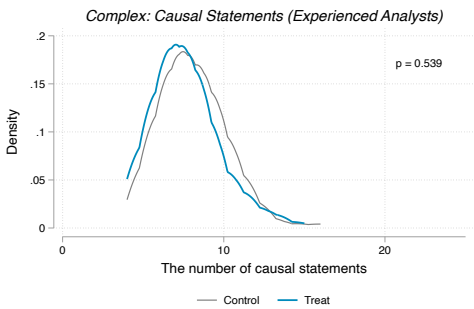
(e)



(f)



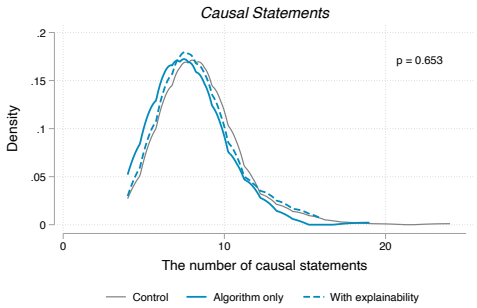
(g)



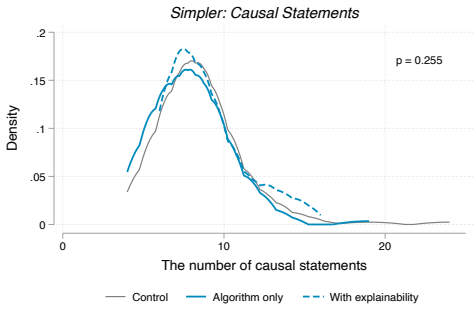
Notes: These figures show kernel density plots of the number of causal statements in analyst reports by treatment and control conditions on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* are included in the plots. * p<0.1, ** p<0.05, *** p<0.01.

Figure A.7: The effects of machine explainability on the number of causal statements by decision type and analyst experience

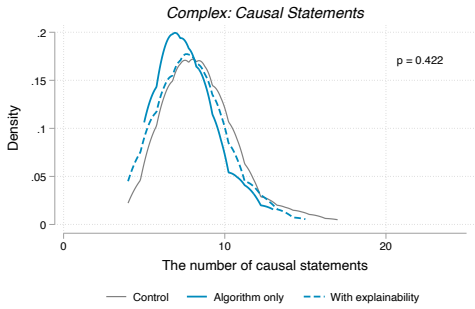
(a)



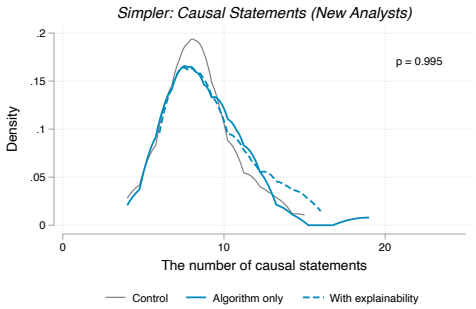
(b)



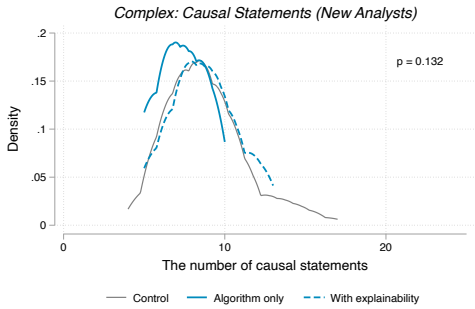
(c)



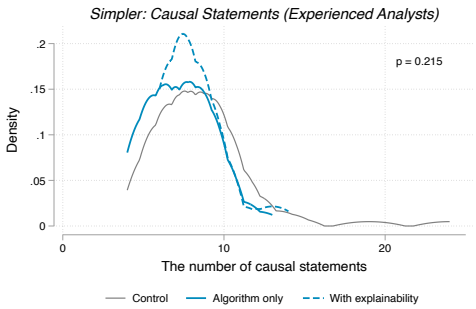
(d)



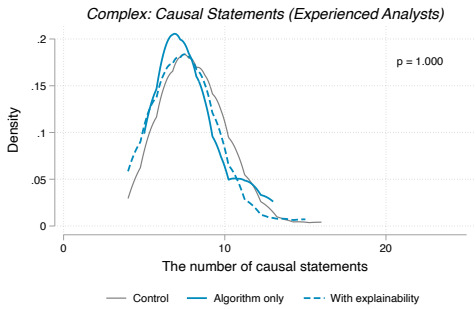
(e)



(f)



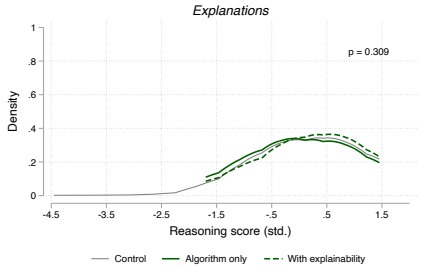
(g)



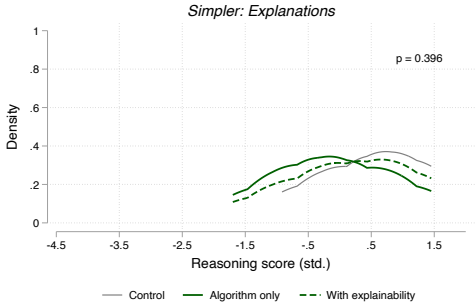
Notes: These figures show kernel density plots of the number of causal statements in analyst reports by treatment condition with algorithm only, treatment condition with additional explainability, and control conditions on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* that compare the two treatment conditions are included in the plots. * p<0.1, ** p<0.05, *** p<0.01.

Figure A.8: The effects of machine explainability on explanations by decision type and analyst experience

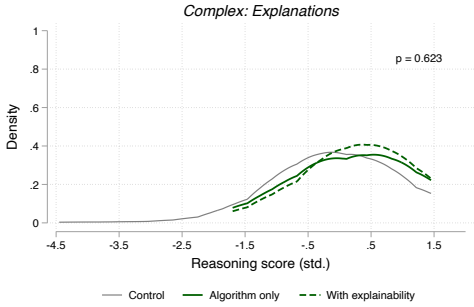
(a)



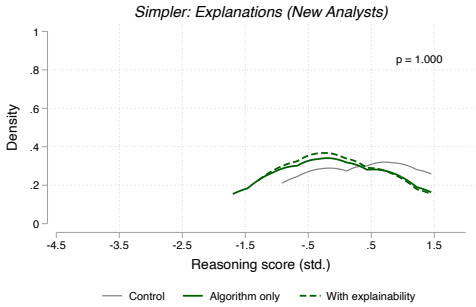
(b)



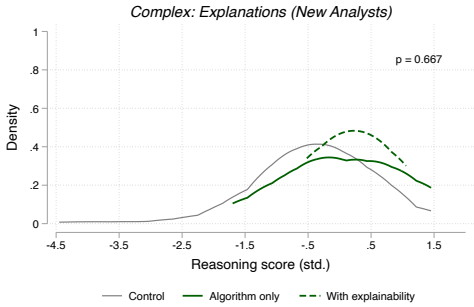
(c)



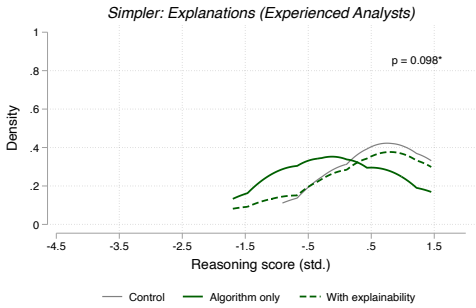
(d)



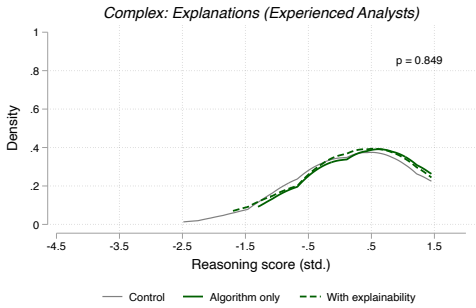
(e)



(f)



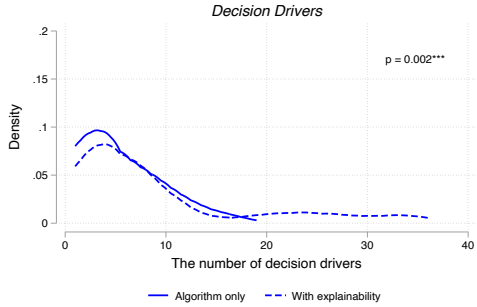
(g)



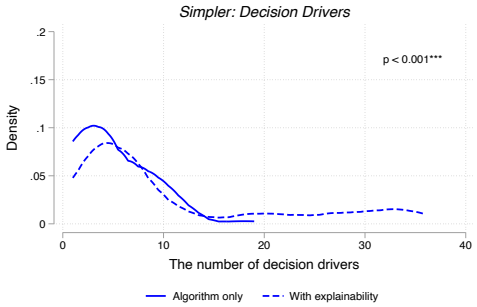
Notes: These figures show kernel density plots of the quality of explanations by treatment condition with machine prediction only, treatment condition with both machine prediction and explainability, and control condition on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* that compare the two treatment conditions are included in the plots. * p<0.1, ** p<0.05, *** p<0.01.

Figure A.9: The effects of machine explainability on the number of decision drivers by decision type and analyst experience

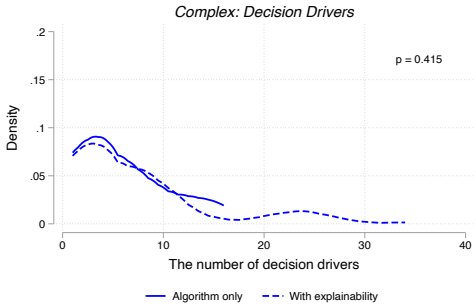
(a)



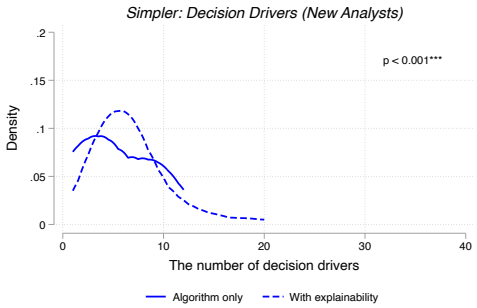
(b)



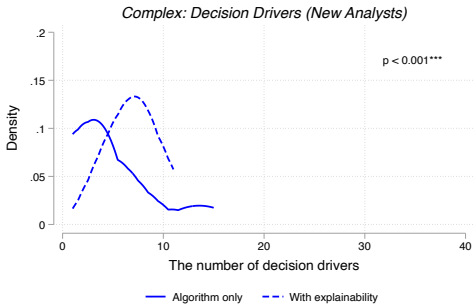
(c)



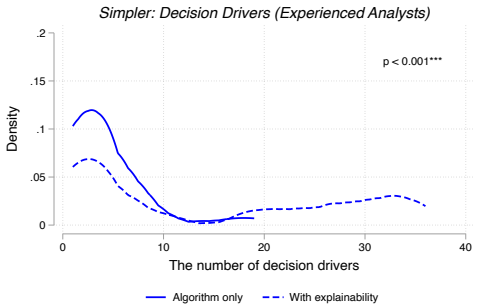
(d)



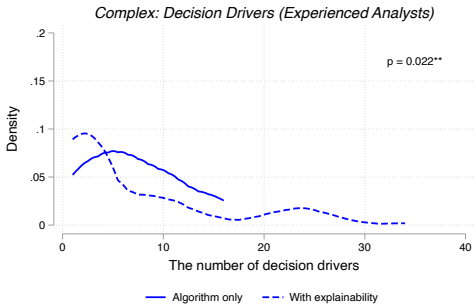
(e)



(f)



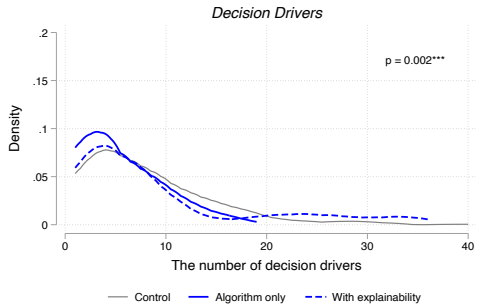
(g)



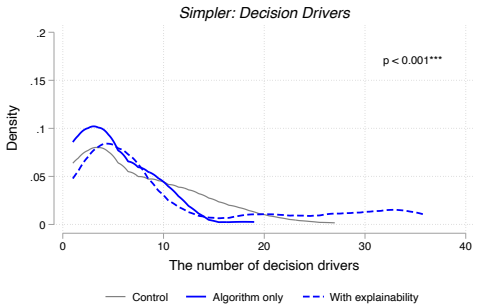
Notes: These figures show kernel density plots of the number of decision drivers treatment condition with machine prediction only and treatment condition with both machine prediction and explainability on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* are included in the plots. * p<0.1, ** p<0.05, *** p<0.01.

Figure A.10: The effects of machine prediction and machine explainability on the number of decision drivers number by decision type and analyst experience

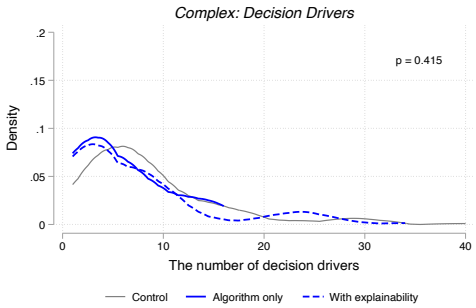
(a)



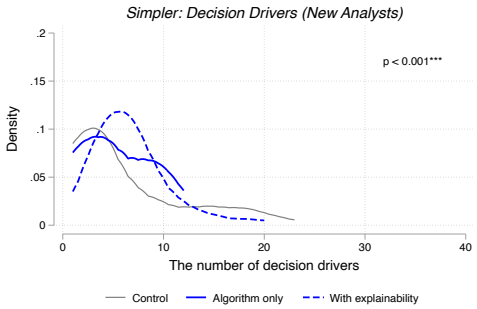
(b)



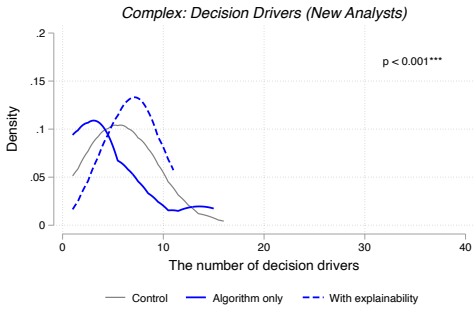
(c)



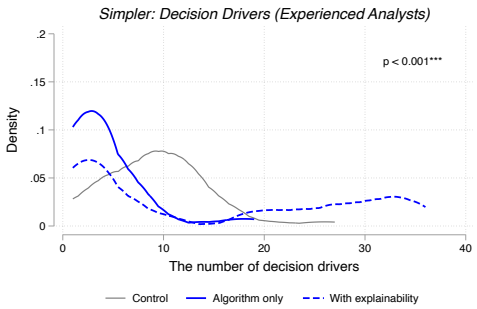
(d)



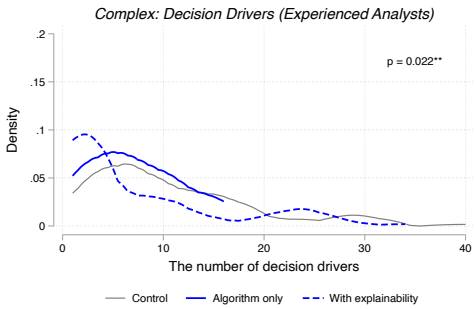
(e)



(f)

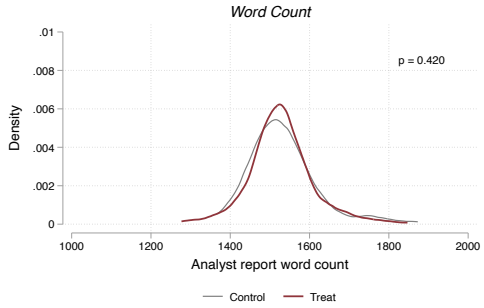


(g)

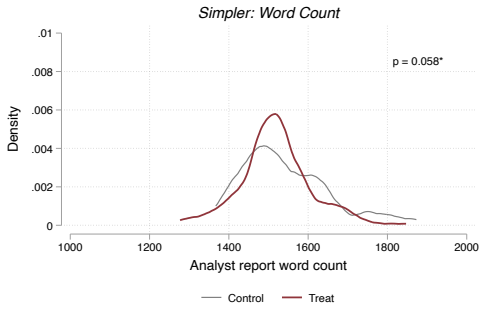


Notes: These figures show kernel density plots of the number of decision drivers by treatment condition with machine prediction only, treatment condition with both machine prediction and explainability, and control condition on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* comparing the two treatment conditions are included in the plots.

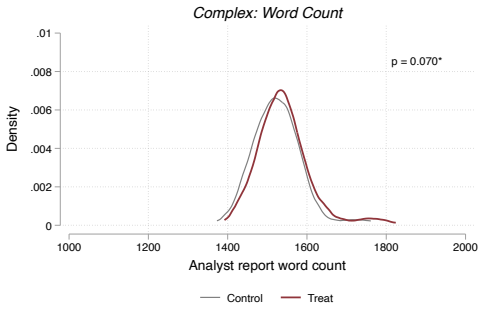
Figure A.11: The effects of machine predictions on analyst report word count by decision type and analyst experience



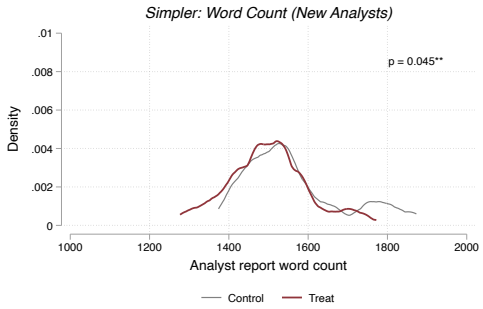
(a)



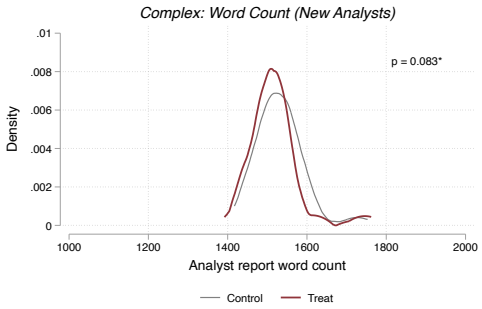
(b)



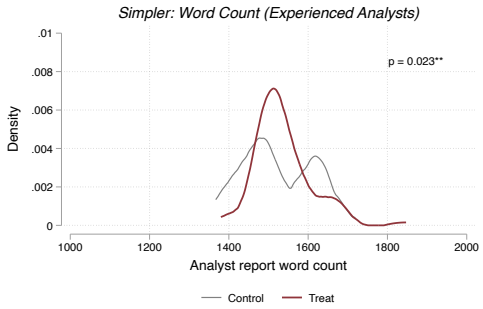
(c)



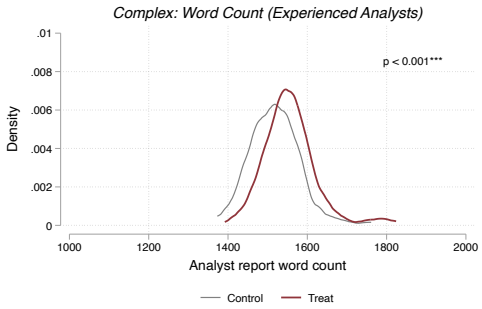
(d)



(e)



(f)

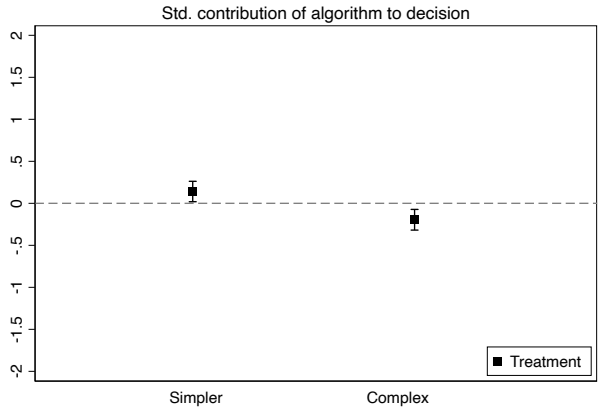


(g)

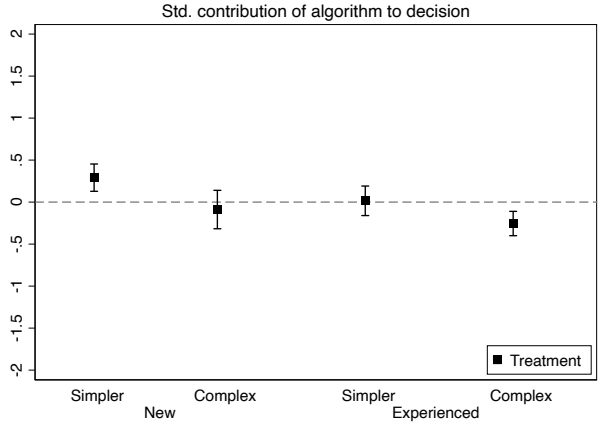
Notes: These figures show kernel density plots of analyst reports word count by treatment and control conditions on average ((a)) and by subgroups ((b)-(g)). Kolmogorov-Smirnov test *p-values* are included in the plots.

Figure A.12: Analyst perceptions of how machine predictions contributed to their decisions (standardized)

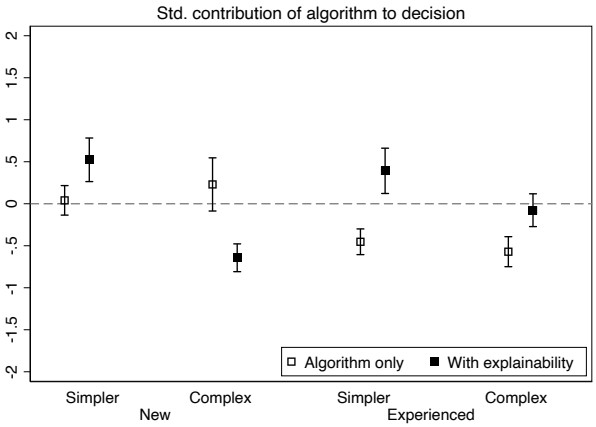
(a) Simpler vs. more complex decisions



(b) Responses by decision type and analyst experience



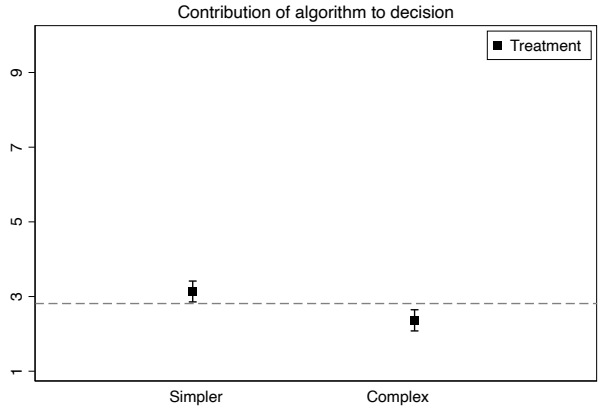
(c) Responses by decision type, analyst experience and machine explainability



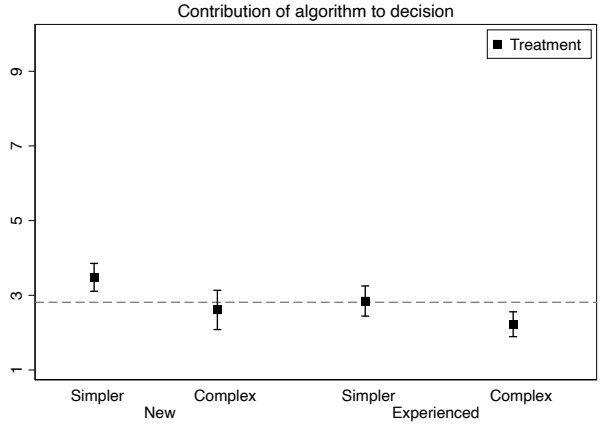
Notes: These figures show standardized means and confidence intervals of analysts' reported contribution of machine predictions to their decisions. Analysts reported the extent to which machine predictions contributed to their rating decisions on a scale between 1 and 10. The ratings reported in this figure are standardized at the analyst level. (a) shows the differential levels of perceived contribution across simpler and complex decisions, (b) additionally presents the differential levels of perceived contribution by analyst experience, and (c) additionally presents the differential levels of perceived contribution by machine explainability. The confidence intervals are calculated at 95% level.

Figure A.13: Analyst perceptions of how machine predictions contributed to their decisions

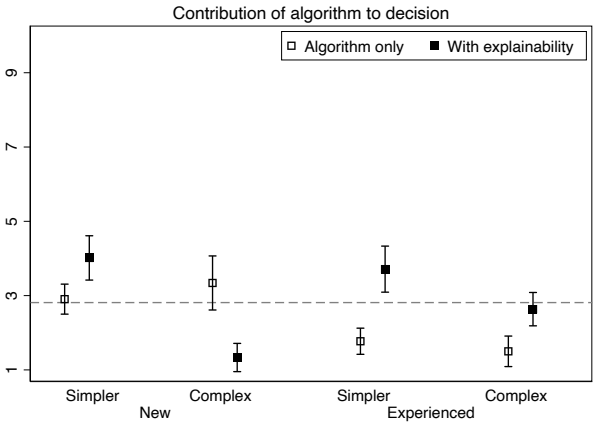
(a) Simpler vs. more complex decisions



(b) Responses by decision type and analyst experience

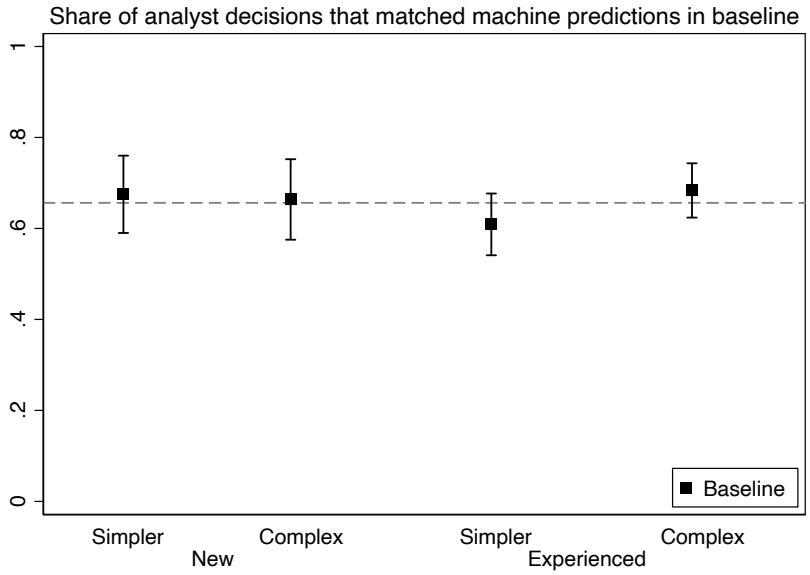


(c) Responses by decision type, analyst experience and machine explainability



Notes: These figures show means and confidence intervals of analysts' reported contribution of machine predictions to their decisions. Analysts reported the extent to which machine predictions contributed to their rating decisions on a scale between 1 and 10. (a) shows the differential levels of perceived contribution across simpler and complex decisions, (b) additionally presents the differential levels of perceived contribution by analyst experience, and (c) additionally presents the differential levels of perceived contribution by machine explainability. The confidence intervals are calculated at 95% level.

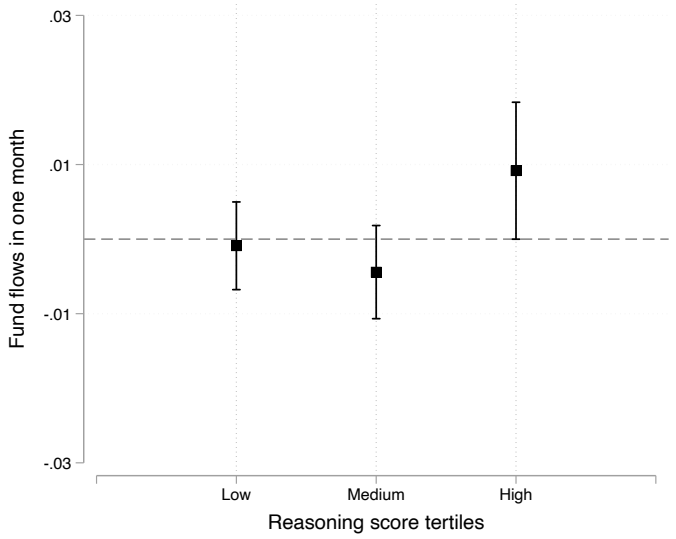
Figure A.14: The share of analyst decisions that matched machine predictions in baseline



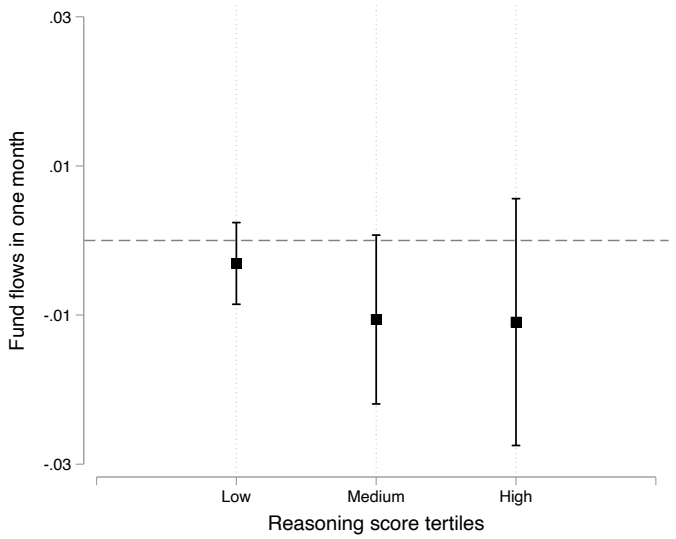
Notes: This figure shows the percentage of analyst decisions that matched machine predictions by decision type, analyst experience in baseline.

Figure A.15: Net fund flows by reasoning scores

(a) Net fund flows of recommended funds by reasoning score tertiles



(b) Net fund flows of not recommended funds by reasoning score tertiles



Notes: These figures show means and confidence intervals of net fund flows in the month after analysts issue fund recommendation. *Net fund flow* measures the difference between the total cash inflows and total cash outflows of a fund in a month, which captures investment activity. A positive net fund flow indicates more money is entering the fund than leaving. (a) shows the differential levels of net fund flows of funds recommended by analysts across reasoning score tertiles, (b) presents the differential levels of fund flows of funds not recommended by analysts across reasoning score tertiles. The confidence intervals are calculated at 95% level.

Table A.1: Separating knowledge and reasoning components of reasoning scores

	Knowledge component score				Logic component score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat × Post	0.004	-0.172	-0.742**	-0.638**	0.021	-0.076	-0.579*	-0.490
	(0.169)	(0.215)	(0.352)	(0.315)	(0.173)	(0.210)	(0.345)	(0.337)
Treat × Post × Complex		0.357	1.020**	0.706*		0.237	0.794**	0.513
		(0.353)	(0.417)	(0.403)		(0.355)	(0.387)	(0.380)
Treat × Post × Experienced			0.853*	0.679*			0.802*	0.656*
			(0.444)	(0.369)			(0.440)	(0.391)
Treat × Post × Complex × Experienced			-1.006	-0.796*			-0.895	-0.796*
			(0.617)	(0.478)			(0.611)	(0.468)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	No	No	No	Yes
Observations	1313	1313	1313	1207	1313	1313	1313	1207
Mean (control)	0.220	0.220	0.220	0.220	0.256	0.256	0.256	0.256
SD (control)	0.975	0.975	0.975	0.975	0.933	0.933	0.933	0.933

Notes: All regressions include a full set of strata and month fixed effects. Regressions in columns (4) and (8) further include review committee fixed effects. The *knowledge component scores* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. The *reasoning component score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.2: Text analysis of analyst reports

	Flesch-Kincaid			Coleman-Liau			Coherence			Word count		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treat × Post	0.052	0.006	-0.475	-0.012	-0.056	-0.463*	-0.005	-0.005	-0.020	-29.273*	-40.091**	-43.299
	(0.153)	(0.242)	(0.399)	(0.118)	(0.177)	(0.259)	(0.005)	(0.007)	(0.014)	(15.070)	(17.768)	(26.702)
Treat × Post × Complex		0.042	0.829		0.066	0.754**		0.009	0.028		-0.711	14.184
		(0.321)	(0.499)		(0.267)	(0.323)		(0.009)	(0.018)		(28.413)	(32.799)
Treat × Post × Experienced			0.841*			0.641*			0.027*			23.788
			(0.501)			(0.336)			(0.015)			(35.630)
Treat × Post × Complex × Experienced			-1.336**			-1.075**			-0.036*			-29.261
			(0.636)			(0.464)			(0.020)			(48.683)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	984	984	984	984	984	984	984	984	984	984	984	984
Mean (control)	11.193	11.193	11.193	13.632	13.632	13.632	0.233	0.233	0.233	1533.126	1533.126	1533.126
SD (control)	1.085	1.085	1.085	0.777	0.777	0.777	0.037	0.037	0.037	86.421	86.421	86.421

Notes: All regressions include a full set of strata and month fixed effects. For all measures except *Word count*, we first processed the text by lemmatizing and removing stop words before constructing the variables. The *Flesch-Kincaid score* in columns (1)-(3) captures the readability of reports based on the number of words in sentences and the average number of syllables per word. The *Coleman-Liau index* in columns (4)-(6) provides another measure of readability based on sentence length and the average number of characters per word. *Coherence* in columns (7)-(9) measures the coherence of topics in the report based on topic modeling. *Word count* in columns (10)-(12) measures the number of words in the analyst report. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.3: The effects of machine explainability on decisions and reasoning

Panel A. Decision change

	Change rating			Reasoning score		
	(1)	(2)	(3)	(4)	(5)	(6)
Explainability × Post	-0.077*	-0.030	0.066	0.192	0.407*	0.175
	(0.044)	(0.048)	(0.062)	(0.171)	(0.228)	(0.287)
Explainability × Post × Complex		-0.115	-0.076		-0.418	0.703**
		(0.098)	(0.145)		(0.349)	(0.327)
Explainability × Post × Experienced			-0.197**			0.578
			(0.087)			(0.404)
Explainability × Post × Complex × Experienced			0.011			-1.533**
			(0.176)			(0.603)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes	Yes
Observations	898	898	898	549	549	549
Mean (control)	0.163	0.163	0.163	0.127	0.127	0.127
SD (control)	0.370	0.370	0.370	0.887	0.887	0.887

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Explainability × Post × Recommend	3.080	0.700	-2.311	-0.328	0.009	-5.087
	(4.422)	(4.987)	(5.423)	(3.358)	(4.089)	(4.491)
Explainability × Post × Recommend × Active Equity		3.345	0.583		-2.083	10.452
		(8.121)	(10.804)		(6.754)	(8.235)
Explainability × Post × Recommend × Experienced			-0.651			-0.146
			(9.743)			(7.361)
Explainability × Post × Recommend × Active Equity × Experienced			8.986			-9.248
			(14.636)			(12.236)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	823	823	823	823	823	823
Mean (control)	-4.467	-4.467	-4.467	-4.832	-4.832	-4.832
SD (control)	13.865	13.865	13.865	11.525	11.525	11.525

Notes: Columns (1)-(3) of Panel A show treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Columns (4)-(6) of Panel A show treatment effects on *reasoning score*, constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund in 6 months after its rating is published. *Explainability* is an indicator variable that takes the value 1 for analysts assigned to the treatment 2 and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise.
Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.4: The effects of machine explainability on decisions and reasoning, separating knowledge and reasoning components

	Knowledge component score				Logic component score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat × Post	-0.009	-0.383	-0.877**	-0.957***	0.019	-0.342	-0.699*	-0.808**
	(0.229)	(0.233)	(0.361)	(0.364)	(0.245)	(0.218)	(0.361)	(0.373)
Explainability × Post	0.018	0.038	-0.632	-0.384	0.023	0.185	-0.476	-0.218
	(0.195)	(0.274)	(0.406)	(0.268)	(0.197)	(0.287)	(0.415)	(0.332)
Treat × Post × Complex		0.735	1.058**	0.851*		0.738	0.903**	0.697
		(0.454)	(0.452)	(0.445)		(0.458)	(0.426)	(0.432)
Explainability × Post × Complex		-0.056	1.245***	1.272***		-0.306	0.723*	0.966**
		(0.386)	(0.403)	(0.318)		(0.380)	(0.400)	(0.415)
Treat × Post × Experienced			0.581	0.840**			0.411	0.714*
			(0.435)	(0.418)			(0.423)	(0.426)
Explainability × Post × Experienced			1.073*	0.517			1.126*	0.570
			(0.550)	(0.377)			(0.573)	(0.435)
Treat × Post × Complex × Experienced			-0.369	-0.710			-0.124	-0.590
			(0.803)	(0.571)			(0.817)	(0.599)
Explainability × Post × Complex × Experienced			-1.791***	-1.524***			-1.540**	-1.533***
			(0.630)	(0.466)			(0.642)	(0.539)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	No	No	No	Yes
Observations	1313	1313	1313	1207	1313	1313	1313	1207
Mean (control)	0.220	0.220	0.220	0.220	0.256	0.256	0.256	0.256
SD (control)	0.975	0.975	0.975	0.975	0.933	0.933	0.933	0.933

	Knowledge component score				Logic component score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Explainability × Post	0.121	0.525*	0.127	0.149	0.095	0.633**	0.221	0.200
	(0.269)	(0.283)	(0.398)	(0.247)	(0.291)	(0.306)	(0.435)	(0.345)
Explainability × Post × Complex		-0.848	0.464	0.716**		-1.110**	-0.026	0.689*
		(0.506)	(0.472)	(0.311)		(0.530)	(0.510)	(0.389)
Explainability × Post × Experienced			0.841	0.558			0.890	0.597
			(0.555)	(0.374)			(0.605)	(0.453)
Explainability × Post × Complex × Experienced			-2.027**	-1.444**			-1.832*	-1.623**
			(0.883)	(0.559)			(0.987)	(0.674)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	No	No	No	Yes
Observations	629	629	629	549	629	629	629	549
Mean (control)	0.104	0.104	0.104	0.104	0.150	0.150	0.150	0.150
SD (control)	0.927	0.927	0.927	0.927	0.899	0.899	0.899	0.899

Notes: The *knowledge component scores* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. The *reasoning component score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. *Treat* is a dummy variable that takes the value 1 for analysts assigned to treatment 1 and 0 otherwise. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.5: The impact of algorithm explainability on analysts' reports based on text analysis

	Flesch-Kincaid			Coleman-Liau			Coherence			Word count		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treat × Post	-0.046	-0.160	-0.577	-0.002	-0.094	-0.351	-0.004	-0.007	-0.032**	-24.079	-31.942*	-6.354
	(0.158)	(0.258)	(0.467)	(0.109)	(0.162)	(0.261)	(0.007)	(0.008)	(0.014)	(16.867)	(18.223)	(25.261)
Explainability × Post	0.177	0.206	-0.410	-0.030	-0.008	-0.537*	-0.007	-0.002	-0.006	-36.869*	-48.790**	-56.692**
	(0.237)	(0.361)	(0.456)	(0.187)	(0.288)	(0.278)	(0.012)	(0.007)	(0.012)	(20.315)	(20.982)	(27.669)
Treat × Post × Complex		0.322	0.754		0.242	0.577*		0.013	0.054***		-17.205	-43.490
		(0.298)	(0.554)		(0.230)	(0.323)		(0.011)	(0.019)		(30.765)	(29.539)
Explainability × Post × Complex		-0.310	0.934*		-0.164	0.950***		-0.011	0.027		14.012	44.041
		(0.485)	(0.532)		(0.396)	(0.331)		(0.033)	(0.025)		(38.270)	(30.191)
Treat × Post × Experienced			0.712			0.383			0.042***			-15.298
			(0.567)			(0.338)			(0.014)			(35.917)
Explainability × Post × Experienced			1.101**			0.915**			0.006			42.263
			(0.526)			(0.353)			(0.015)			(33.222)
Treat × Post × Complex × Experienced			-0.774			-0.606			-0.072***			33.172
			(0.681)			(0.442)			(0.021)			(51.800)
Explainability × Post × Complex × Experienced			-2.030***			-1.667***			-0.063			-64.348
			(0.690)			(0.506)			(0.047)			(52.801)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	984	984	984	984	984	984	986	986	986	984	984	984
Mean (control)	11.193	11.193	11.193	13.632	13.632	13.632	0.233	0.233	0.233	1533.126	1533.126	1533.126
SD (control)	1.085	1.085	1.085	0.777	0.777	0.777	0.037	0.037	0.037	86.421	86.421	86.421

Notes: All regressions include a full set of strata and month fixed effects. For all measures except Word count, we first processed the text by lemmatizing and removing stop words before constructing the variables. The *Flesch-Kincaid score* in columns (1)-(3) captures the readability of reports based on the number of words in sentences and the average number of syllables per word. The *Coleman-Liau index* in columns (4)-(6) provides another measure of readability based on sentence length and the average number of characters per word. *Coherence* in columns (7)-(9) measures the coherence of topics in the report based on topic modeling. Word count in columns (10)-(12) measures the number of words in the analyst report. *Treat* is a dummy variable that takes the value 1 for analysts assigned to treatment 1 and 0 otherwise. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise.

Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.6: The impact of algorithmic predictions on standardized stated confidence

	Confidence before interview			Confidence after interview		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.247**	0.375*	0.256	0.147	0.296	0.101
	(0.124)	(0.205)	(0.165)	(0.136)	(0.222)	(0.189)
Treat × Post × Complex		-0.315	-0.578**		-0.394	-0.451
		(0.257)	(0.265)		(0.278)	(0.371)
Treat × Post × Experienced			0.256			0.360
			(0.372)			(0.400)
Treat × Post × Complex × Experienced			0.229			-0.033
			(0.466)			(0.543)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1887	1887	1887
Mean (control)	-0.071	-0.071	-0.071	-0.041	-0.041	-0.041
SD (control)	0.885	0.885	0.885	0.920	0.920	0.920

	Confidence before interview			Confidence after interview		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.121	0.193	0.231	0.060	0.143	0.102
	(0.111)	(0.150)	(0.167)	(0.113)	(0.150)	(0.176)
Explainability × Post	0.394*	0.549	0.161	0.263	0.463	0.065
	(0.212)	(0.370)	(0.174)	(0.235)	(0.401)	(0.279)
Treat × Post × Complex		-0.161	-0.401		-0.219	-0.261
		(0.222)	(0.280)		(0.240)	(0.369)
Explainability × Post × Complex		-0.428	-0.677**		-0.577	-0.666
		(0.433)	(0.284)		(0.457)	(0.427)
Treat × Post × Experienced			-0.142			-0.026
			(0.259)			(0.264)
Explainability × Post × Experienced			0.728			0.743
			(0.613)			(0.697)
Treat × Post × Complex × Experienced			0.310			0.035
			(0.404)			(0.462)
Explainability × Post × Complex × Experienced			0.037			-0.120
			(0.709)			(0.809)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1887	1887	1887
Mean (control)	-0.071	-0.071	-0.071	-0.041	-0.041	-0.041
SD (control)	0.885	0.885	0.885	0.920	0.920	0.920

Notes: The *confidence before interview* is the standardized stated confidence in their decisions reported by analysts on a scale of 1-10 before they conduct interview and additional research. The *confidence after interview* the standardized stated confidence in their decisions reported by analysts on a scale of 1-10 before they conduct interview and additional research. *Treat* is a dummy variable that takes the value 1 for analysts assigned to treatment 1 and 0 otherwise. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.7: The impact of machine predictions on decisions and reasoning

	Change rating			Reasoning score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.080**	0.062	0.093**	-0.103	-0.059	-0.054
	(0.031)	(0.045)	(0.046)	(0.118)	(0.154)	(0.153)
Treat × Post × High Baseline Confidence		0.056			-0.003	
		(0.066)			(0.213)	
Treat × Post × High Baseline Reasoning Score			-0.030			-0.090
			(0.066)			(0.235)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes	Yes
Observations	1780	1780	1780	1207	1207	1207
Mean (control)	0.113	0.113	0.113	0.238	0.238	0.238
SD (control)	0.317	0.317	0.317	0.928	0.928	0.928

Notes: All regressions include a full set of strata and month fixed effects.
Regressions in columns (4)-(6) further include review committee fixed effects. Columns (1)-(3) show treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Columns (4)-(6) show treatment effects on *reasoning score*, constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *High baseline confidence* takes the value 1 if the analyst has an average stated confidence after interview that is above average at baseline and 0 otherwise. *High baseline reasoning score* takes the value 1 if the analyst has an average reasoning score that is above average at baseline and 0 otherwise.

Table A.8: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating		Reasoning score	
	(1)	(2)	(3)	(4)
Treat × Post	0.080**	0.094***	-0.103	-0.161
	(0.031)	(0.033)	(0.118)	(0.118)
Treat × Post × New coverage		-0.133***		0.531
		(0.048)		(0.418)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Committee FE	No	No	Yes	Yes
Observations	1780	1780	1207	1207
Mean (control)	0.113	0.113	0.238	0.238
SD (control)	0.317	0.317	0.928	0.928

Panel B. Decision performance

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	0.894	-0.070	1.130	0.555
	(3.033)	(2.954)	(2.165)	(2.284)
Treat × Post × Recommend × New Coverage		15.054		12.369
		(11.808)		(8.223)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	1670	1670	1668	1668
Mean (control)	-4.467	-4.467	-4.832	-4.832
SD (control)	13.865	13.865	11.525	11.525

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund in 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *New coverage* takes the value 1 if it is the first time that an analyst covers a given fund and 0 otherwise. All regressions include a full set of strata and month fixed effects, and regressions in Panel B additionally control for fund age and size.
Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.9: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating		Reasoning score	
	(1)	(2)	(3)	(4)
Treat × Post	0.080**	0.166***	-0.103	0.197
	(0.031)	(0.049)	(0.118)	(0.244)
Treat × Post × Male		-0.115*		-0.401
		(0.065)		(0.286)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Committee FE	No	No	Yes	Yes
Observations	1780	1780	1207	1207
Mean (control)	0.113	0.113	0.238	0.238
SD (control)	0.317	0.317	0.928	0.928

Panel B. Decision performance

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	0.894	1.203	1.130	5.424
	(3.033)	(6.518)	(2.165)	(3.463)
Treat × Post × Recommend × Male		-0.025		-5.239
		(7.387)		(4.313)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	1670	1670	1668	1668
Mean (control)	-4.467	-4.467	-4.832	-4.832
SD (control)	13.865	13.865	11.525	11.525

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund in 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *Male* takes the value 1 if the analyst is male and 0 if the analyst is female. All regressions include a full set of strata and month fixed effects, and regressions in Panel B additionally control for fund age and size.
Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.10: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating		Reasoning score	
	(1)	(2)	(3)	(4)
Treat × Post	0.080**	0.080**	-0.103	-0.071
	(0.031)	(0.032)	(0.118)	(0.122)
Treat × Post × Manager		-0.024		-0.525
		(0.101)		(0.433)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Committee FE	No	No	Yes	Yes
Observations	1780	1780	1207	1207
Mean (control)	0.113	0.113	0.238	0.238
SD (control)	0.317	0.317	0.928	0.928

Panel B. Decision performance

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	0.894	1.748	1.130	1.338
	(3.033)	(3.219)	(2.165)	(2.285)
Treat × Post × Recommend × Manager		-15.320**		-4.750
		(6.379)		(4.432)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	1670	1670	1668	1668
Mean (control)	-4.467	-4.467	-4.832	-4.832
SD (control)	13.865	13.865	11.525	11.525

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund in 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *Male* takes the value 1 if the analyst is male and 0 otherwise. All regressions include a full set of strata and month fixed effects, and regressions in Panel B additionally control for fund age and size.
Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.11: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating		Reasoning score	
	(1)	(2)	(3)	(4)
Treat × Post	0.080**	0.088**	-0.103	-0.120
	(0.031)	(0.044)	(0.118)	(0.173)
Treat × Post × Postgraduate		-0.010		0.025
		(0.065)		(0.227)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Committee FE	No	No	Yes	Yes
Observations	1780	1780	1207	1207
Mean (control)	0.113	0.113	0.238	0.238
SD (control)	0.317	0.317	0.928	0.928

Panel B. Decision performance

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	0.894	-5.856	1.130	-3.132
	(3.033)	(3.598)	(2.165)	(2.372)
Treat × Post × Recommend × Postgraduate		15.359**		8.961*
		(6.443)		(4.728)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	1670	1670	1668	1668
Mean (control)	-4.467	-4.467	-4.832	-4.832
SD (control)	13.865	13.865	11.525	11.525

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund in 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *Postgraduate* takes the value 1 if the analyst has a post-graduate degree and 0 otherwise. All regressions include a full set of strata and month fixed effects, and regressions in Panel B additionally control for fund age and size.
Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.12: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating					
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.074**	0.075**	0.080**	0.080**	0.080*	0.067
	(0.033)	(0.033)	(0.031)	(0.031)	(0.043)	(0.060)
Treat × Post × Complex					-0.000	0.007
					(0.071)	(0.106)
Treat × Post × Experienced						0.033
						(0.090)
Treat × Post × Complex × Experienced						-0.029
						(0.144)
Strata FE	No	Yes	No	Yes	Yes	Yes
Month FE	No	No	Yes	Yes	Yes	Yes
Observations	1780	1780	1780	1780	1780	1780
Mean (control)	0.113	0.113	0.113	0.113	0.113	0.113
SD (control)	0.317	0.317	0.317	0.317	0.317	0.317

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.894	9.886***	13.499***	1.130	7.949***	8.885**
	(3.033)	(2.585)	(3.440)	(2.165)	(2.487)	(3.620)
Treat × Post × Recommend × Active Equity		-16.822***	-14.568*		-11.458***	-8.163
		(5.268)	(7.764)		(4.191)	(6.830)
Treat × Post × Recommend × Experienced			-9.966*			-3.784
			(5.374)			(4.691)
Treat × Post × Recommend × Active Equity × Experienced			-0.198			-1.867
			(11.332)			(9.125)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.467	-4.467	-4.467	-4.832	-4.832	-4.832
SD (control)	13.865	13.865	13.865	11.525	11.525	11.525

Notes: Panel A shows treatment effects on Change rating, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on Fund returns in 3 months, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on Fund returns in 6 months, the risk-adjusted return of a fund in 6 months after its rating is published. Treatment is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. Post takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than the medium rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include a full set of strata and month fixed effects, and regressions in Panel B additionally control for fund age and size.
Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.13: The impact of machine predictions on reasoning

	All committees				Single
	(1)	(2)	(3)	(4)	(5)
Treat × Post	0.012	-0.124	-0.292	-0.366	-0.423
	(0.169)	(0.210)	(0.307)	(0.249)	(0.277)
Treat × Post × Complex		0.297	0.440	0.301	0.707*
		(0.351)	(0.364)	(0.299)	(0.376)
Treat × Post × Experienced			0.410	0.520	0.533
			(0.445)	(0.333)	(0.373)
Treat × Post × Complex × Experienced			-0.478	-0.537	-0.821
			(0.645)	(0.438)	(0.526)
Strata FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes
Observations	1313	1313	1313	1207	901
Mean (control)	0.238	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.928	0.968

Notes: All regressions include a full set of strata and month fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of committees with one reviewer only. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than the medium rating experience at the company and 0 otherwise.
Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.14: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating				
	(1)	(2)	(3)	(4)	(5)
Treat × Post	0.074**	0.075**	0.080**	0.080**	
	(0.033)	(0.033)	(0.031)	(0.031)	
Treat x Post x New x Simpler					0.076
					(0.046)
Treat x Post x New x Complex					0.045
					(0.097)
Treat x Post x Experienced x Simpler					0.079
					(0.054)
Treat x Post x Experienced x Complex					0.100*
					(0.057)
Strata FE	No	Yes	No	Yes	Yes
Month FE	No	No	Yes	Yes	Yes
Observations	1780	1780	1780	1780	1780
Mean (control)	0.113	0.113	0.113	0.113	0.113
SD (control)	0.317	0.317	0.317	0.317	0.317

Panel B. Decision performance

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	0.894		1.130	
	(3.033)		(2.165)	
Treat x Post x Recommend x New x Simpler		9.933***		6.359
		(3.439)		(4.085)
Treat x Post x Recommend x New x Complex		-5.204		0.718
		(6.912)		(4.783)
Treat x Post x Recommend x Experienced x Simpler		6.418*		6.077**
		(3.536)		(2.879)
Treat x Post x Recommend x Experienced x Complex		-6.328		-4.788
		(5.267)		(3.973)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	1670	1670	1668	1668
Mean (control)	-4.467	-4.467	-4.832	-4.832
SD (control)	13.865	13.865	11.525	11.525

Notes: Panel A shows treatment effects on Change rating, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on Fund returns in 3 months, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on Fund returns in 6 months, the risk-adjusted return of a fund in 6 months after its rating is published. Treatment is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. Post takes the value 1 after the treatment is implemented and 0 otherwise. Complex takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. Experienced takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Recommend takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include a full set of strata and month fixed effects, and regressions in Panel B additionally control for fund age and size.
Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.15: The impact of machine predictions on reasoning

	All committees			Single
	(1)	(2)	(3)	(4)
Treat × Post	0.012			
	(0.169)			
Treat x Post x New x Simpler		-0.509*	-0.368*	-0.231
		(0.261)	(0.221)	(0.213)
Treat x Post x New x Complex		0.660**	0.366*	0.545**
		(0.259)	(0.215)	(0.247)
Treat x Post x Experienced x Simpler		-0.047	-0.018	-0.040
		(0.249)	(0.194)	(0.201)
Treat x Post x Experienced x Complex		-0.024	-0.295	-0.152
		(0.352)	(0.187)	(0.224)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Committee FE	No	No	Yes	Yes
Observations	1313	1313	1207	901
Mean (control)	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.968

Notes: All regressions include a full set of strata and month fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of committees with one reviewer only. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.16: The number of decision drivers and causal statements by experimental condition

	Decision drivers		Causal statements	
	(1)	(2)	(3)	(4)
Treat × Post	-0.332*		-0.065	
	(0.187)		(0.040)	
Treat x Post x New x Simpler		-0.047		-0.016
		(0.348)		(0.050)
Treat x Post x New x Complex		-0.547		-0.184***
		(0.409)		(0.052)
Treat x Post x Experienced x Simpler		-0.492		-0.157**
		(0.404)		(0.068)
Treat x Post x Experienced x Complex		-0.487		0.079
		(0.295)		(0.063)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	1317	1317	948	948
Mean (control)	1.728	1.728	2.104	2.104
SD (control)	0.896	0.896	0.274	0.274

Notes: All regressions include a full set of strata and month fixed effects. *Decision drivers* measure the number of reported decision drivers by the analyst and is natural log transformed. *Causal statements* measure the number of causal statements in the analyst report as identified by GPT-4 and is natural log transformed. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. Post takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise.
Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.17: The impact of algorithm explainability on decisions and causal explanations

Panel A: Decision change				
	Change rating		Reasoning score	
	(1)	(2)	(3)	(4)
Treat × Post	0.120***		0.005	
	(0.041)		(0.235)	
Explainability × Post	0.040		0.020	
	(0.034)		(0.194)	
Treat x Post x New x Simpler		0.038		-0.602**
		(0.056)		(0.253)
Treat x Post x New x Complex		0.059		0.603**
		(0.130)		(0.291)
Treat x Post x Experienced x Simpler		0.146**		-0.483**
		(0.063)		(0.230)
Treat x Post x Experienced x Complex		0.224***		0.295
		(0.074)		(0.616)
Explainability x Post x New x Simpler		0.107*		-0.411
		(0.055)		(0.357)
Explainability x Post x New x Complex		0.098*		0.832***
		(0.057)		(0.270)
Explainability x Post x Experienced x Simpler		0.007		0.290
		(0.060)		(0.361)
Explainability x Post x Experienced x Complex		0.018		-0.283
		(0.045)		(0.280)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	1780	1780	1313	1313
Panel B: Decision performance				
	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post	-1.278		-1.941	
	(2.072)		(2.296)	
Explainability × Post	-3.100		-0.092	
	(2.120)		(2.694)	
Treat × Post × Recommend	0.002	-9.388*	1.834	-1.865
	(4.065)	(5.120)	(2.547)	(3.704)
Explainability × Post × Recommend	2.235	-3.169	0.322	-6.295
	(3.357)	(6.759)	(3.118)	(5.915)
Treat x Post x New x Simpler		-1.605		1.247
		(2.614)		(3.696)
Treat x Post x New x Complex		4.428		-2.080
		(5.208)		(7.220)
Treat x Post x Experienced x Simpler		-2.542		-2.996
		(1.958)		(1.879)
Treat x Post x Experienced x Complex		-3.432		-3.822
		(2.679)		(3.001)
Explainability x Post x New x Simpler		-5.419		5.917
		(3.958)		(5.600)
Explainability x Post x New x Complex		-2.238		-9.630**
		(7.288)		(4.086)
Explainability x Post x Experienced x Simpler		-5.452*		-4.355**
		(2.870)		(2.100)
Explainability x Post x Experienced x Complex		0.285		3.651
		(4.472)		(5.070)
Treat x Post x Recommend x New x Simpler		23.282***		11.422*
		(6.431)		(5.931)

Table A.18: The depth and breadth of decision drivers and causal statements by experimental condition

	Breadth of decision drivers			Depth of decision drivers		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	-0.386	-0.027	0.134	-0.331	0.257	1.239**
	(0.251)	(0.336)	(0.433)	(0.448)	(0.556)	(0.600)
Treat × Post × Complex		-0.613	-0.936		-0.922	-2.524*
		(0.444)	(0.616)		(0.867)	(1.308)
Treat × Post × Experienced			-0.135			-1.659
			(0.574)			(1.065)
Treat × Post × Complex × Experienced			0.263			2.540
			(0.779)			(1.709)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1317	1317	1317	1317	1317	1317
Mean (control)	2.596	2.596	2.596	4.003	4.003	4.003
SD (control)	1.216	1.216	1.216	2.404	2.404	2.404

Notes: All regressions include a full set of strata and month fixed effects. *Breadth of decision drivers* measures to what extend the reported decision drivers by the analyst span across five major categories of decision drivers. *Depth of decision drivers* measures depth of decision drivers by counting the number of decision drivers in the category where an analyst reports the largest number of decision drivers. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. Post takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.19: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating		
	(1)	(2)	(3)
Treat × Post	0.075**	0.070	0.109*
	(0.033)	(0.045)	(0.057)
Treat × Post × Complex		0.006	-0.047
		(0.072)	(0.119)
Treat × Post × Experienced			-0.045
			(0.086)
Treat × Post × Complex × Experienced			0.060
			(0.149)
Strata FE	Yes	Yes	Yes
Observations	1780	1780	1780
Mean (control)	0.113	0.113	0.113
SD (control)	0.317	0.317	0.317

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	1.247	12.367***	14.468***	1.123	8.543***	6.704*
	(3.178)	(2.654)	(3.365)	(2.250)	(2.563)	(3.679)
Treat × Post × Recommend × Active Equity		-21.020***	-19.903***		-12.566***	-8.136
		(5.161)	(7.204)		(4.313)	(6.443)
Treat × Post × Recommend × Experienced			-7.326			0.148
			(4.892)			(4.396)
Treat × Post × Recommend × Active Equity × Experienc			4.800			-1.575
			(10.114)			(8.415)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.467	-4.467	-4.467	-4.832	-4.832	-4.832
SD (control)	13.865	13.865	13.865	11.525	11.525	11.525

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund in 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include strata fixed effects, and regressions in Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.20: The impact of machine predictions on reasoning

	All committees				Single
	(1)	(2)	(3)	(4)	(5)
Treat × Post	0.007	-0.131	-0.663*	-0.542*	-0.661*
	(0.166)	(0.211)	(0.343)	(0.317)	(0.333)
Treat × Post × Complex		0.301	0.910**	0.599	1.124**
		(0.349)	(0.398)	(0.384)	(0.445)
Treat × Post × Experienced			0.814*	0.652*	0.758*
			(0.435)	(0.370)	(0.393)
Treat × Post × Complex × Experienced			-0.939	-0.799*	-1.300**
			(0.603)	(0.462)	(0.530)
Strata FE	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes
Observations	1313	1313	1313	1207	901
Mean (control)	0.238	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.928	0.968

Notes: All regressions include strata fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of committees with one reviewer only. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.21: The number of decision drivers and causal statements by experimental condition

	Decision drivers			Causal statements		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	-0.319*	0.040	0.443	-0.061	-0.103*	0.041
	(0.191)	(0.260)	(0.320)	(0.041)	(0.053)	(0.083)
Treat × Post × Complex		-0.609*	-1.374**		0.097	-0.200*
		(0.352)	(0.543)		(0.080)	(0.115)
Treat × Post × Experienced			-0.776*			-0.243**
			(0.459)			(0.107)
Treat × Post × Complex × Experienced			1.266*			0.481***
			(0.666)			(0.147)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1317	1317	1317	948	948	948
Mean (control)	1.728	1.728	1.728	2.104	2.104	2.104
SD (control)	0.896	0.896	0.896	0.274	0.274	0.274

Notes: All regressions include strata fixed effects. *Decision drivers* measure the number of reported decision drivers by the analyst and is natural log transformed. *Causal statements* measure the number of causal statements in the analyst report as identified by GPT-4 and is natural log transformed. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. Post takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.22: The impact of algorithm explainability on decisions and reasoning

Panel A: Decision change						
	Change rating			Reasoning score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.111**	0.082	0.070	0.004	-0.359	-0.812**
	(0.043)	(0.051)	(0.067)	(0.233)	(0.225)	(0.361)
Explainability × Post	0.039	0.056	0.150***	0.010	0.094	-0.540
	(0.037)	(0.051)	(0.057)	(0.189)	(0.277)	(0.395)
Treat × Post × Complex		0.067	-0.007		0.727	1.009**
		(0.102)	(0.150)		(0.449)	(0.440)
Explainability × Post × Complex		-0.045	-0.006		-0.162	0.954**
		(0.068)	(0.084)		(0.385)	(0.391)
Treat × Post × Experienced			0.060			0.553
			(0.097)			(0.425)
Explainability × Post × Experienced			-0.153			1.023*
			(0.093)			(0.552)
Treat × Post × Complex × Experienced			0.087			-0.323
			(0.185)			(0.789)
Explainability × Post × Complex × Experienced			-0.000			-1.565**
			(0.122)			(0.632)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1780	1780	1780	1313	1313	1313
Panel B: Decision performance						
	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.418	13.524***	18.246***	2.187	10.252***	10.589**
	(4.193)	(3.436)	(3.948)	(2.655)	(3.175)	(4.391)
Explainability × Post × Recommend	2.524	12.003***	14.319***	-0.123	6.919**	2.976
	(3.684)	(3.737)	(5.247)	(3.118)	(3.279)	(4.755)
Treat × Post × Recommend × Active Equity		-24.363***	-22.685***		-13.090**	-12.380
		(5.966)	(8.499)		(5.051)	(7.527)
Explainability × Post × Recommend × Complex		-18.022**	-23.613***		-11.938**	-3.070
		(6.931)	(7.945)		(6.002)	(7.170)
Treat × Post × Recommend × Experienced			-10.135*			-2.229
			(5.604)			(5.178)
Explainability × Post × Recommend × Experienced			-11.585			-1.177
			(8.956)			(6.589)
Treat × Post × Recommend × Active Equity × Experienc			1.595			3.223
			(10.856)			(9.114)
Explainability × Post × Recommend × Complex ×			17.415			-2.408
			(13.136)			(10.804)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668

Notes: *Treat* is an indicator variable that takes the value 1 for analysts assigned to treatment 1 who receive only the algorithmic prediction. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 who receive algorithmic predictions along with key fund features that contributed most to the prediction based on Shapley values. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include strata fixed effects, and regressions in Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.23: The impact of machine predictions on decisions

Panel A. Decision change

	Change rating		
	(1)	(2)	(3)
Treat × Post	0.080**	0.074*	0.105*
	(0.031)	(0.044)	(0.059)
Treat × Post × Complex		0.003	-0.034
		(0.071)	(0.126)
Treat × Post × Experienced			-0.035
			(0.086)
Treat × Post × Complex × Experienced			0.039
			(0.154)
Month FE	Yes	Yes	Yes
Observations	1780	1780	1780
Mean (control)	0.113	0.113	0.113
SD (control)	0.317	0.317	0.317

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	1.183	10.044***	10.746***	1.094	7.943***	5.693
	(3.101)	(2.777)	(3.820)	(2.239)	(2.441)	(3.968)
Treat × Post × Recommend × Complex		-17.437***	-16.506**		-12.119***	-7.769
		(5.344)	(7.706)		(4.083)	(6.303)
Treat × Post × Recommend × Experienced			-3.264			2.721
			(5.430)			(4.939)
Treat × Post × Recommend × Complex × Experienced			3.338			-2.799
			(10.535)			(8.445)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.467	-4.467	-4.467	-4.832	-4.832	-4.832
SD (control)	13.865	13.865	13.865	11.525	11.525	11.525

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund in 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include month fixed effects, and regressions in Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.24: The impact of machine predictions on reasoning

	All committees				Single
	(1)	(2)	(3)	(4)	(5)
Treat × Post	-0.002	-0.055	-0.613*	-0.510*	-0.603*
	(0.177)	(0.218)	(0.316)	(0.292)	(0.318)
Treat × Post × Complex		0.188	0.785**	0.528	1.005**
		(0.354)	(0.358)	(0.350)	(0.407)
Treat × Post × Experienced			0.818*	0.658*	0.726*
			(0.418)	(0.352)	(0.381)
Treat × Post × Complex × Experienced			-0.885	-0.778*	-1.220**
			(0.590)	(0.437)	(0.498)
Month FE	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes
Observations	1313	1313	1313	1207	901
Mean (control)	0.238	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.928	0.968

Notes: All regressions include month fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of committees with one reviewer only. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.25: The number of decision drivers and causal statements by experimental condition

	Decision drivers			Causal statements		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	-0.319	0.082	0.476	-0.067	-0.123**	0.033
	(0.199)	(0.297)	(0.330)	(0.043)	(0.055)	(0.075)
Treat × Post × Complex		-0.718*	-1.419**		0.142*	-0.201*
		(0.386)	(0.552)		(0.084)	(0.107)
Treat × Post × Experienced			-0.873			-0.252**
			(0.547)			(0.099)
Treat × Post × Complex × Experienced			1.330*			0.539***
			(0.730)			(0.140)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1317	1317	1317	948	948	948
Mean (control)	1.728	1.728	1.728	2.104	2.104	2.104
SD (control)	0.896	0.896	0.896	0.274	0.274	0.274

Notes: All regressions include month fixed effects. *Decision drivers* measure the number of reported decision drivers by the analyst and is natural log transformed. *Causal statements* measure the number of causal statements in the analyst report as identified by GPT-4 and is natural log transformed. *Treatment* is a dummy variable that takes the value 1 for analysts assigned to treatment and 0 otherwise. Post takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at analyst level. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table A.26: The impact of algorithm explainability on decisions and reasoning

Panel A: Decision change						
	Change rating			Reasoning score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.119***	0.092*	0.072	-0.085	-0.383	-0.854**
	(0.040)	(0.049)	(0.067)	(0.245)	(0.244)	(0.355)
Explainability × Post	0.040	0.055	0.138**	0.084	0.259	-0.394
	(0.034)	(0.051)	(0.065)	(0.197)	(0.277)	(0.352)
Treat × Post × Complex		0.058	0.012		0.681	0.958**
		(0.099)	(0.158)		(0.460)	(0.414)
Explainability × Post × Complex		-0.046	-0.024		-0.323	0.810**
		(0.069)	(0.096)		(0.371)	(0.336)
Treat × Post × Experienced			0.070			0.602
			(0.094)			(0.448)
Explainability × Post × Experienced			-0.143			0.979*
			(0.097)			(0.501)
Treat × Post × Complex × Experienced			0.041			-0.289
			(0.192)			(0.806)
Explainability × Post × Complex × Experienced			0.023			-1.544***
			(0.129)			(0.582)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1780	1780	1780	1313	1313	1313
Panel B: Decision performance						
	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.098	10.143***	14.597***	1.436	7.988**	8.540*
	(4.094)	(3.746)	(4.758)	(2.644)	(3.221)	(4.729)
Explainability × Post × Recommend	2.739	10.658***	10.634*	0.711	7.911**	2.127
	(3.361)	(3.659)	(5.384)	(2.961)	(3.008)	(5.122)
Treat × Post × Recommend × Active Equity		-19.162***	-19.567**		-11.025**	-11.302
		(6.465)	(9.712)		(4.842)	(7.280)
Explainability × Post × Recommend × Complex		-15.794**	-19.994***		-12.782**	-1.779
		(6.773)	(6.870)		(5.735)	(7.809)
Treat × Post × Recommend × Experienced			-9.257			-1.349
			(6.376)			(5.863)
Explainability × Post × Recommend × Experienced			-0.672			7.629
			(8.116)			(6.457)
Treat × Post × Recommend × Active Equity × Experienc			5.594			4.785
			(11.913)			(9.261)
Explainability × Post × Recommend × Complex ×			7.538			-11.419
			(11.782)			(10.914)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668

Notes: *Treat* is an indicator variable that takes the value 1 for analysts assigned to treatment 1 who receive only the algorithmic prediction. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 who receive algorithmic predictions along with key fund features that contributed most to the prediction based on Shapley values. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include month fixed effects, and regressions in Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at analyst level. * p<0.05, ** p<0.01, *** p<0.001.

Table A.27: Share of complex decisions by analyst tenure

	Pre-treatment	Post-treatment	Full sample
	(1)	(2)	(3)
Experienced	-0.006	0.028	0.013
	(0.038)	(0.029)	(0.023)
Strata FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
Observations	672	1215	1887

Notes: All regressions include a full set of strata and month fixed effects. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Share complex* measures the percentage of complex funds an analyst rated in a given month. Regressions in columns (1) include the subsample of baseline data. Columns (2) include the subsample of post-experiment data. Columns (3) include the full sample. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.28: Impact of algorithmic predictions on reasoning score by the share of complex decisions

	Analysts who cover simpler funds only		Subsample of simpler funds		Full sample	
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	-0.125	-0.647*	-0.121	-0.612	-0.137	-0.665*
	(0.219)	(0.379)	(0.219)	(0.394)	(0.216)	(0.353)
Treat × Post × Experienced		0.774		0.716		0.808*
		(0.468)		(0.479)		(0.451)
Treat × Post × Share Complex			1.045	1.337	0.310	0.943**
			(0.975)	(1.958)	(0.387)	(0.440)
Treat × Post × Share Complex × Experienced				-0.373		-0.954
				(2.253)		(0.677)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	504	504	605	605	1313	1313
Mean (control)	0.549	0.549	0.535	0.535	0.238	0.238
SD (control)	0.784	0.784	0.824	0.824	0.928	0.928

Notes: All regressions include a full set of strata and month fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Share complex* measures the percentage of complex funds an analyst rated in a given month. Regressions in columns (1)-(2) include the subsample of analysts who exclusively covered simpler funds in a month. Columns (3)-(4) include the subsample of simpler funds covered by analysts. Columns (5)-(6) include the full sample. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.29: Compare the returns of the portfolios constructed based on funds recommended by treated analysts with the portfolios constructed based on funds recommended by controlled analysts

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	-2.109	4.744	11.652**	0.090	7.317**	13.508**
	(2.878)	(3.264)	(5.843)	(1.925)	(2.894)	(5.957)
Treat × Post × Share Complex		-11.460**	-15.090		-11.710***	-20.552***
		(5.551)	(10.565)		(3.572)	(7.763)
Treat × Post × Experienced			-14.181**			-12.511*
			(6.818)			(6.404)
Treat × Post × Share Complex ×			9.969			16.553*
Experienced			(12.275)			(8.580)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	524	524	524	524	524	524
Mean (control)	-4.586	-4.586	-4.586	-5.504	-5.504	-5.504
SD (control)	12.388	12.388	12.388	10.289	10.289	10.289

Notes: All regressions include a full set of strata and month fixed effects. The analysis constructs and compares two hypothetical investment portfolios, one based on treated analyst's rating decision, and the other based on controlled analyst's rating decision. *Treat* takes the value 1 if the portfolio is constructed based on treated analyst's decision and 0 otherwise. *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Share complex* measures the percentage of complex funds an analyst rated in a given month. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.30: Compare the returns of the portfolios constructed based on funds recommended by controlled analysts with the portfolios constructed based on funds recommended by machine predictions

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Analyst Rated × Post	0.203	1.022	2.115	0.059	0.630	0.919
	(0.859)	(1.207)	(2.571)	(0.496)	(0.686)	(1.316)
Analyst Rated × Post × Share Complex		-1.599	-3.641		-1.209	-1.509
		(1.721)	(3.391)		(1.032)	(1.364)
Analyst Rated × Post × Experienced			-1.486			-0.309
			(2.957)			(1.575)
Analyst Rated × Post × Share Complex ×			2.964			0.393
Experienced			(3.869)			(1.861)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	534	534	534	534	534	534
Mean (control)	-4.627	-4.627	-4.627	-5.502	-5.502	-5.502
SD (control)	12.485	12.485	12.485	10.217	10.217	10.217

Notes: All regressions include a full set of strata and month fixed effects. Only analysts in the control group are included with the analysis, whose recommendation decisions are compared with machine recommendations of the same funds. The analysis constructs and compares two hypothetical investment portfolios, one based on treated analyst's rating decision, and the other based on machine prediction. *Analyst Rated* takes the value 1 if the portfolio is constructed based on analyst's decision and 0 otherwise. *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Share complex* measures the percentage of complex funds an analyst rated in a given month. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.31: Compare the returns of the portfolios constructed based on funds recommended by treated analysts with the portfolios constructed based on funds recommended by machine predictions

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Analyst Rated × Post	0.666	0.892	0.317	1.167**	1.211	0.081
	(0.821)	(1.369)	(1.522)	(0.530)	(0.891)	(0.905)
Analyst Rated × Post × Share Complex		-0.214	0.944		0.030	1.466
		(1.645)	(2.714)		(1.155)	(1.966)
Analyst Rated × Post × Experienced			-0.938			0.519
			(2.447)			(1.345)
Analyst Rated × Post × Share Complex × Experienced			0.035			-1.074
			(3.129)			(2.362)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	528	528	528	528	528	528
Mean (control)	-4.610	-4.610	-4.610	-5.871	-5.871	-5.871
SD (control)	13.619	13.619	13.619	9.998	9.998	9.998

Notes: All regressions include a full set of strata and month fixed effects. Only analysts in the treatment group are included with the analysis, whose recommendation decisions are compared with machine recommendations of the same funds. The analysis constructs and compares two hypothetical investment portfolios, one based on treated analyst's rating decision, and the other based on machine prediction. *Analyst Rated* takes the value 1 if the portfolio is constructed based on analyst's decision and 0 otherwise. *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Share complex* measures the percentage of complex funds an analyst rated in a given month. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.32: The impact of algorithmic predictions on the duration of time spent on research platform to rate a fund

	Duration on research platform (mins)			Days spent on research platform		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.891*	1.361**	1.626*	0.451	7.666**	15.549*
	(0.500)	(0.559)	(0.871)	(2.276)	(2.951)	(8.460)
Treat × Post × Complex		-0.606	-1.954		-14.318***	-16.722
		(1.054)	(1.426)		(4.380)	(10.154)
Treat × Post × Experienced			-0.453			-11.233
			(1.119)			(8.853)
Treat × Post × Complex × Experienced			1.812			2.799
			(1.996)			(10.920)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1819	1819	1819	1819	1819	1819
Mean (control)	1.390	1.390	1.390	4.266	4.266	4.266
SD (control)	2.131	2.131	2.131	13.984	13.984	13.984

Notes: All regressions include a full set of strata and month fixed effects. The table reports the impact of algorithmic predictions on the duration of time spent on research platform to rate a fund. *Duration on research platform (mins)* measures the number of minutes an analyst logged on the research platform. *Days spent on research platform* counts the number of days an analyst between the first time an analyst research a fund on the platform and when the analyst submits their decision. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Share complex* measures the percentage of complex funds an analyst rated in a given month. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.33: The impact of algorithmic predictions on the duration of time spent on research platform to rate a fund

	Duration on research platform (mins)			Days spent on research platform		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.705	1.607***	2.015**	1.282	6.222*	15.401*
	(0.518)	(0.578)	(0.920)	(3.122)	(3.189)	(8.627)
Explainability × Post	1.059	1.107*	1.201	-0.138	9.059***	15.723*
	(0.750)	(0.641)	(0.937)	(2.542)	(2.972)	(8.459)
Treat × Post × Complex		-1.609	-2.027		-8.833	-11.073
		(1.090)	(1.764)		(6.010)	(10.187)
Explainability × Post × Complex		0.434	-2.212*		-19.311***	-24.253**
		(1.533)	(1.128)		(4.767)	(10.397)
Treat × Post × Experienced			-0.728			-13.712
			(1.179)			(9.122)
Treat × Post × Complex × Experienced			0.592			-0.855
			(2.144)			(11.339)
Explainability × Post × Experienced			-0.124			-8.842
			(1.271)			(8.925)
Explainability × Post × Complex ×			2.997			8.001
Experienced			(2.193)			(11.611)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1819	1819	1819	1819	1819	1819
Mean (control)	1.390	1.390	1.390	4.266	4.266	4.266
SD (control)	2.131	2.131	2.131	13.984	13.984	13.984

Notes: All regressions include a full set of strata and month fixed effects. The table reports the impact of algorithmic predictions on the duration of time spent on research platform to rate a fund. *Duration on research platform (mins)* measures the number of minutes an analyst logged on the research platform. *Days spent on research platform* counts the number of days an analyst between the first time an analyst research a fund on the platform and when the analyst submits their decision. *Treat* is an indicator variable that takes the value 1 for analysts assigned to treatment 1 who receive only the algorithmic prediction. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 who receive algorithmic predictions along with key fund features that contributed most to the prediction based on Shapley values. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Share complex* measures the percentage of complex funds an analyst rated in a given month. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.34: Robustness check on the impact of machine predictions on decisions with analyst fixed effects

Panel A. Decision change

	Change rating					
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.074**	0.085**	0.073**	0.084**	0.087*	0.110***
	(0.033)	(0.035)	(0.032)	(0.035)	(0.045)	(0.039)
Treat × Post × Complex					-0.003	-0.034
					(0.075)	(0.129)
Treat × Post × Experienced						-0.026
						(0.078)
Treat × Post × Complex × Experienced						0.031
						(0.161)
Analyst FE	No	Yes	No	Yes	Yes	Yes
Month FE	No	No	Yes	Yes	Yes	Yes
Observations	1780	1779	1780	1779	1779	1779
Mean (control)	0.113	0.113	0.113	0.113	0.113	0.113
SD (control)	0.317	0.317	0.317	0.317	0.317	0.317

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.543	9.092***	11.055***	1.715	7.042***	5.568*
	(3.185)	(2.383)	(3.248)	(2.131)	(2.173)	(3.014)
Treat × Post × Recommend × Active		-16.959***	-21.790**		-9.889**	-13.128**
Equity		(5.357)	(9.135)		(3.942)	(6.057)
Treat × Post × Recommend × Experienced			-4.480			0.080
			(4.612)			(3.801)
Treat × Post × Recommend × Active			10.531			7.267
Equity × Experienced			(11.424)			(7.833)
Analyst FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1667	1667	1667	1665	1665	1665
Mean (control)	-4.188	-4.188	-4.188	-4.243	-4.243	-4.243
SD (control)	13.620	13.620	13.620	12.812	12.812	12.812

Notes: All regressions include a full set of analyst and month fixed effects. Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.35: Robustness check on the impact of machine predictions on reasoning with analyst fixed effects

	All committees				Single
	(1)	(2)	(3)	(4)	(5)
Treat × Post	-0.066	-0.019	-0.305	-0.156	-0.201
	(0.145)	(0.170)	(0.298)	(0.195)	(0.207)
Treat × Post × Complex		-0.005	0.409	0.062	0.613
		(0.274)	(0.381)	(0.254)	(0.337)
Treat × Post × Experienced			0.509	0.227	0.281
			(0.352)	(0.238)	(0.249)
Treat × Post × Complex × Experienced			-0.674	-0.159	-0.561
			(0.507)	(0.326)	(0.403)
Analyst FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes
Observations	1311	1311	1311	1205	899
Mean (control)	0.238	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.928	0.968

Notes: All regressions include a full set of analyst and month fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of committees with one reviewer only. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.36: Robust check on the number of decision drivers and causal statements by experimental condition with analyst fixed effects

	Decision Drivers			Causal Statements		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.017	-0.119	-0.152	-0.057	-0.128**	0.021
	(0.110)	(0.145)	(0.188)	(0.043)	(0.055)	(0.066)
Treat × Post × Active Equity		0.205	0.377		0.135	-0.184*
		(0.189)	(0.245)		(0.083)	(0.094)
Treat × Post × Experienced			0.060			-0.237**
			(0.279)			(0.095)
Treat × Post × Active Equity ×			-0.303			0.510***
Experienced			(0.357)			(0.133)
Analyst FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1312	1312	1312	940	940	940
Mean (control)	1.728	1.728	1.728	2.104	2.104	2.104
SD (control)	0.896	0.896	0.896	0.274	0.274	0.274

Notes: All regressions include a full set of analyst and month fixed effects. *Decision drivers* measure the number of reported decision drivers by the analyst and is natural log transformed. *Causal statements* measure the number of causal statements in the analyst report as identified by GPT-4 and is natural log transformed. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.37: Robustness check on the impact of machine predictions on decisions with analyst fixed effects

Panel A. Decision change

	Change rating			Reasoning score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.119**	0.107**	0.080	-0.050	-0.137	-0.325
	(0.046)	(0.051)	(0.052)	(0.199)	(0.147)	(0.259)
Explainability × Post	0.050	0.067	0.138***	-0.082	0.081	-0.263
	(0.036)	(0.050)	(0.041)	(0.172)	(0.244)	(0.437)
Treat × Post × Complex		0.031	-0.012		0.249	0.389
		(0.105)	(0.155)		(0.372)	(0.386)
Explainability × Post × Complex		-0.037	-0.025		-0.253	0.557
		(0.070)	(0.086)		(0.287)	(0.440)
Treat × Post × Experienced			0.064			0.303
			(0.088)			(0.317)
Treat × Post × Complex × Experienced			0.054			-0.233
			(0.194)			(0.739)
Explainability × Post × Experienced			-0.123			0.627
			(0.091)			(0.497)
Explainability × Post × Complex ×			0.020			-1.118**
Experienced			(0.129)			(0.531)
Analyst FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1779	1779	1779	1311	1311	1311
Mean (control)	0.113	0.113	0.113	0.238	0.238	0.238
SD (control)	0.317	0.317	0.317	0.928	0.928	0.928

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	-0.663	9.469***	14.387***	1.965	7.632***	8.053**
	(4.265)	(2.849)	(3.675)	(2.606)	(2.517)	(3.457)
Explainability × Post × Recommend	2.367	8.848**	10.944**	1.469	5.944*	1.883
	(3.578)	(3.378)	(4.871)	(2.974)	(3.530)	(4.490)
Treat × Post × Recommend × Active		-20.300***	-23.060**		-11.025**	-15.938**
Equity		(6.129)	(10.695)		(4.270)	(6.769)
Explainability × Post × Recommend ×		-12.598*	-32.350***		-7.720	-10.335*
Active Equity		(6.751)	(7.157)		(5.786)	(5.755)
Treat × Post × Recommend × Experienced			-9.239*			-1.282
			(4.943)			(4.288)
Treat × Post × Recommend × Active			8.275			9.430
Equity × Experienced			(12.640)			(8.257)
Explainability × Post × Recommend ×			-2.693			0.288
Experienced			(7.397)			(6.465)
Explainability × Post × Recommend ×			24.439**			8.082
Active Equity × Experienced			(11.214)			(9.357)
Analyst FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1667	1667	1667	1665	1665	1665
Mean (control)	-4.188	-4.188	-4.188	-4.243	-4.243	-4.243
SD (control)	13.620	13.620	13.620	12.812	12.812	12.812

Notes: Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. , and on *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. . Panel B shows whether recommended funds by analysts in different treatment conditions observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Treat* is an indicator variable that takes the value 1 for analysts assigned to treatment 1 who receive only the algorithmic prediction. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 who re-

Table A.38: The impact of machine predictions on reasoning by analyst experience and decision type

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.894	2.824	11.557***	1.130	-0.161	6.741*
	(3.033)	(3.843)	(3.600)	(2.165)	(3.007)	(3.761)
Treat × Post × Recommend × Experienced		-4.051	-5.959		1.750	-0.604
		(5.483)	(5.031)		(4.051)	(4.510)
Treat × Post × Recommend × Complex			-16.747**			-8.323
			(7.857)			(6.471)
Treat × Post × Recommend × Complex			5.281			-0.092
× Experienced			(10.532)			(8.418)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.188	-4.188	-4.188	-4.243	-4.243	-4.243
SD (control)	13.620	13.620	13.620	12.812	12.812	12.812

Notes: The table shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. Column (2) and (5) show the alternative model specification to those in Table 1B, with analyst experience as the moderator. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include a full set of strata and month fixed effects. Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at the analyst level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A.39: The impact of machine predictions on reasoning by analyst experience and decision type

	All committees				Single
	(1)	(2)	(3)	(4)	(5)
Treat × Post	0.012	-0.337	-0.661*	-0.564*	-0.698**
	(0.169)	(0.209)	(0.345)	(0.322)	(0.339)
Treat × Post × Experienced		0.485	0.828*	0.668*	0.786*
		(0.306)	(0.438)	(0.374)	(0.398)
Treat × Post × Complex			0.907**	0.609	1.151**
			(0.395)	(0.387)	(0.446)
Treat × Post × Complex × Experienced			-0.950	-0.796*	-1.293**
			(0.607)	(0.465)	(0.529)
Strata FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes
Observations	1313	1313	1313	1207	901
Mean (control)	0.238	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.928	0.968

Notes: All regressions include a full set of strata and month fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of committees with one reviewer only. Column (2) shows the alternative model specification to that in Table 2, with analyst experience as the moderator. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.40: Heterogeneous treatment effect by whether the fund was outside the analyst's specialization

	Change rating		Fund return		Reasoning score	
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.096**		-0.905		-0.018	
	(0.041)		(2.062)		(0.216)	
Treat × Post × Outside analyst's specialization	-0.055		-5.066		0.124	
	(0.072)		(3.390)		(0.382)	
Treat × Post		0.134**		-0.265		0.088
		(0.055)		(2.567)		(0.315)
Treat × Post × Outside analyst's specialization		-0.047		-4.433		-0.196
		(0.087)		(3.562)		(0.496)
Explainability × Post		0.058		-1.763		-0.133
		(0.043)		(2.582)		(0.226)
Explainability × Post × Outside analyst's specialization		-0.057		-6.017		0.456
		(0.082)		(5.076)		(0.386)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1780	1780	1670	1670	1313	1313
Mean (control)	0.113	0.113	-4.188	-4.188	0.238	0.238
SD (control)	0.317	0.317	13.620	13.620	0.928	0.928

Notes:

Table A.41: The impact of algorithm explainability on analyst reports

	Number of Shapley features discussed			Percentage of Shapley features discussed		
	(1)	(2)	(3)	(4)	(5)	(6)
Explainability × Post	0.003	-0.119	1.157***	0.169	-0.192	5.705***
	(0.306)	(0.465)	(0.227)	(1.417)	(2.115)	(1.479)
Explainability × Post × Complex		0.485	-0.849		1.858	-4.522
		(0.585)	(0.619)		(2.776)	(3.098)
Explainability × Post × Experienced			-1.993***			-9.365***
			(0.592)			(2.912)
Explainability × Post × Complex × Experienced			1.824*			8.763*
			(0.901)			(4.379)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	466	466	466	466	466	466
Mean (control)	2.796	2.796	2.796	14.051	14.051	14.051
SD (control)	1.211	1.211	1.211	6.044	6.044	6.044

Notes: All regressions include a full set of strata and month fixed effects. *Number of Shapley features discussed* counts the number of Shapley features that are discussed in the analyst reports. *Percentage of Shapley features discussed* measures the percentage of Shapley features that are discussed in the analyst reports. *Explainability* is a dummy variable that takes the value 1 for analysts assigned to treatment 2 who receive algorithmic predictions along with key fund features that contributed most to the prediction based on Shapley values. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.42: Robustness check on the impact of machine predictions on decisions, measuring experience as a continuous variable

Panel A. Decision change

	Change rating					
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post	0.074**	0.075**	0.080**	0.080**	0.080*	0.064
	(0.033)	(0.033)	(0.031)	(0.031)	(0.043)	(0.064)
Treat × Post × Complex					-0.000	0.017
					(0.071)	(0.109)
Treat × Post × Tenure						0.004
						(0.007)
Treat × Post × Complex × Tenure						-0.003
						(0.010)
Strata FE	No	Yes	No	Yes	Yes	Yes
Month FE	No	No	Yes	Yes	Yes	Yes
_cons	Yes	No	No	No	No	No
Observations	1780	1780	1780	1780	1780	1780
Mean (control)	0.113	0.113	0.113	0.113	0.113	0.113
SD (control)	0.317	0.317	0.317	0.317	0.317	0.317

Panel B. Decision performance

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.894	9.886***	15.103***	1.130	7.949***	9.284**
	(3.033)	(2.585)	(4.175)	(2.165)	(2.487)	(4.252)
Treat × Post × Recommend × Complex		-16.822***	-22.684***		-11.458***	-12.024*
		(5.268)	(7.813)		(4.191)	(6.675)
Treat × Post × Recommend × Tenure			-0.802*			-0.198
			(0.435)			(0.411)
Treat × Post × Recommend × Complex			0.900			0.216
× Tenure			(0.558)			(0.556)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.188	-4.188	-4.188	-4.243	-4.243	-4.243
SD (control)	13.620	13.620	13.620	12.812	12.812	12.812

Notes: All regressions include a full set of strata and month fixed effects. Panel A shows treatment effects on *Change rating*, which is an indicator variable that takes the value 1 if the rating decision by the analyst differs from the currently assigned rating. Panel B shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) of Panel B show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) of Panel B show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Tenure* is a continuous measure of analyst experience, calculated as the number of years of analyst tenure at the firm. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. Panel B additionally control for fund age and size. Standard errors are in parentheses, clustered at the analyst level. Random inference p-values are in square brackets, calculated using 5000 repetitions. Standard errors are clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.43: Robustness check on the impact of machine predictions on explanations, measuring experience as a continuous variable

	All committees				Single
	(1)	(2)	(3)	(4)	(5)
Treat × Post	0.012	-0.124	-0.437	-0.566*	-0.691*
	(0.169)	(0.210)	(0.359)	(0.314)	(0.356)
Treat × Post × Complex		0.297	0.972**	0.733**	1.294***
		(0.351)	(0.478)	(0.366)	(0.450)
Treat × Post × Tenure			0.043	0.066*	0.081*
			(0.043)	(0.037)	(0.044)
Treat × Post × Complex × Tenure			-0.083*	-0.094**	-0.143***
			(0.048)	(0.039)	(0.050)
Strata FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
Committee FE	No	No	No	Yes	Yes
Observations	1313	1313	1313	1207	901
Mean (control)	0.238	0.238	0.238	0.238	0.286
SD (control)	0.928	0.928	0.928	0.928	0.968

Notes: All regressions include a full set of strata and month fixed effects. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include reasoning scores issued by committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of reasoning scores issued by committees with one reviewer only. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Tenure* is a continuous measure of analyst experience, calculated as the number of years of analyst tenure at the firm. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.44: Sample counts by treatment groups

	Control	Algorithm only	With explainability
Number of analysts	49	24	24
Number of new analysts	16	11	7
Number of experienced analysts	33	13	17
Number of new analysts specializing in equity funds	13	5	7
Number of new analysts specializing in non-equity funds	12	6	5
Number of exp. analysts specializing in equity funds	10	6	5
Number of exp. analysts specializing in non-equity funds	14	7	7

Notes: This table reports the number of analysts, the number of analysts by experience and specialization, and by control group, treatment group with algorithmic predictions only, and treatment group with both algorithmic predictions and explainability, respectively.

Table A.45: Robustness check on the impact of machine predictions on decisions, reporting fund control variables

	Fund returns in 3 months			Fund returns in 6 months		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat × Post × Recommend	0.894	9.886***	11.557***	1.130	7.949***	6.741*
	(3.033)	(2.585)	(3.600)	(2.165)	(2.487)	(3.761)
Treat × Post × Recommend × Complex		-16.822***	-16.747**		-11.458***	-8.323
		(5.268)	(7.857)		(4.191)	(6.471)
Treat × Post × Recommend × Experienced			-5.959			-0.604
			(5.031)			(4.510)
Treat × Post × Recommend × Complex			5.281			-0.092
× Experienced			(10.532)			(8.418)
Fund Age	0.040	0.032	0.025	0.069***	0.060**	0.055**
	(0.032)	(0.032)	(0.032)	(0.024)	(0.023)	(0.023)
Fund Size	0.088***	0.088***	0.083***	0.074***	0.069***	0.063**
	(0.032)	(0.030)	(0.029)	(0.026)	(0.026)	(0.026)
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1670	1670	1670	1668	1668	1668
Mean (control)	-4.188	-4.188	-4.188	-4.243	-4.243	-4.243
SD (control)	13.620	13.620	13.620	12.812	12.812	12.812

Notes: This table is a robustness check of Table 3 Panel B, which additionally presents the coefficients of control variables, fund age and size. *Fund age* is the number of years since the inception of the fund. *Fund size* is approximated by the total net assets under management (in billions of USD). The table shows whether recommended funds by treated analysts observe higher fund returns in subsequent months. Columns (1)-(3) show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. All regressions include a full set of strata and month fixed effects. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.46: Impact of machine predictions on subsample of simpler decisions

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	9.496***	11.587***	7.772***	6.715*
	(2.487)	(3.756)	(2.628)	(3.997)
Treat × Post × Recommend × Experienced		-6.366		-0.549
		(4.969)		(4.625)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	838	838	837	837
Mean (control)	-5.125	-5.125	-6.267	-6.267
SD (control)	12.854	12.854	14.433	14.433

Notes: All regressions include a full set of strata and month fixed effects. The table presents the impact of machine predictions on the subsample of simpler decisions only. Columns (1)-(3) show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.47: Impact of machine predictions on subsample of complex decisions

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	-7.403	-7.290	-4.488	-3.288
	(4.602)	(6.023)	(3.151)	(3.862)
Treat × Post × Recommend × Experienced		1.952		0.455
		(8.660)		(6.089)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	830	830	829	829
Mean (control)	-3.216	-3.216	-2.142	-2.142
SD (control)	14.329	14.329	10.496	10.496

Notes: All regressions include a full set of strata and month fixed effects. The table presents the impact of machine predictions on the subsample of complex decisions only. Columns (1)-(3) show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.48: Impact of machine predictions on subsample of new analysts

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	3.655	12.031***	0.021	6.377
	(3.933)	(3.870)	(3.066)	(3.997)
Treat × Post × Recommend × Complex		-17.257**		-7.455
		(8.415)		(7.192)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	722	722	722	722
Mean (control)	-3.355	-3.355	-2.716	-2.716
SD (control)	14.804	14.804	15.314	15.314

Notes: All regressions include a full set of strata and month fixed effects. The table presents the impact of machine predictions on the subsample of new analysts only. Columns (1)-(3) show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.49: Impact of machine predictions on subsample of experienced analysts

	Fund returns in 3 months		Fund returns in 6 months	
	(1)	(2)	(3)	(4)
Treat × Post × Recommend	-1.559	5.773	1.319	6.165**
	(3.863)	(3.473)	(2.649)	(2.553)
Treat × Post × Recommend × Complex		-11.842		-8.437
		(7.132)		(5.369)
Strata FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	948	948	946	946
Mean (control)	-4.782	-4.782	-5.339	-5.339
SD (control)	12.695	12.695	10.548	10.548

Notes: All regressions include a full set of strata and month fixed effects. The table presents the impact of machine predictions on the subsample of experienced analysts only. Columns (1)-(3) show treatment effects on *Fund returns in 3 months*, the risk-adjusted return of a fund 3 months after its rating is published. Columns (4)-(6) show treatment effects on *Fund returns in 6 months*, the risk-adjusted return of a fund 6 months after its rating is published. *Recommend* takes the value 1 for funds that analysts recommend and 0 otherwise. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.50: The impact of machine predictions on explanations on subsample of simpler decisions

	All committees		Single
	(1)	(2)	(3)
Treat × Post	-0.083	-0.578*	-0.588
	(0.213)	(0.340)	(0.373)
Treat × Post × Experienced		0.718	0.710*
		(0.437)	(0.419)
Strata FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
Committee FE	No	No	Yes
Observations	605	605	456
Mean (control)	0.535	0.535	0.533
SD (control)	0.824	0.824	0.854

Notes: All regressions include a full set of strata and month fixed effects. The table presents the impact of machine predictions on the subsample of simpler decisions only. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include reasoning scores issued by committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of reasoning scores issued by committees with one reviewer only. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.51: The impact of machine predictions on explanations on subsample of complex decisions

	All committees		Single
	(1)	(2)	(3)
Treat × Post	0.173	0.228	0.445
	(0.279)	(0.203)	(0.291)
Treat × Post × Experienced		-0.113	-0.434
		(0.418)	(0.358)
Strata FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
Committee FE	No	No	Yes
Observations	706	706	441
Mean (control)	-0.006	-0.006	0.039
SD (control)	0.938	0.938	1.014

Notes: All regressions include a full set of strata and month fixed effects. The table presents the impact of machine predictions on the subsample of complex decisions only. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include reasoning scores issued by committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of reasoning scores issued by committees with one reviewer only. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Experienced* takes the value 1 if the analyst has more than three years of rating experience at the company and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * p<0.1, ** p<0.05, *** p<0.01.

Table A.52: The impact of machine predictions on explanations on subsample of new analysts

	All committees		Single
	(1)	(2)	(3)
Treat × Post	-0.294	-0.632*	-0.451
	(0.215)	(0.352)	(0.336)
Treat × Post × Complex		0.921**	1.046**
		(0.404)	(0.451)
Strata FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
Committee FE	No	No	Yes
Observations	532	532	363
Mean (control)	-0.003	-0.003	0.047
SD (control)	0.994	0.994	1.038

Notes: All regressions include a full set of strata and month fixed effects. The table presents the impact of machine predictions on the subsample of new analysts only. Regressions in columns (4)-(5) further include review committee fixed effects. *Reasoning score* is constructed at the committee level by standardizing individual scores and taking an average across all reviewers in the committee. Regressions in columns (1)-(4) include reasoning scores issued by committees of all sizes, which vary between 1 reviewer and 4 reviewers per committee. Column (5) analyzes the subsample of reasoning scores issued by committees with one reviewer only. *Treatment* is an indicator variable that takes the value 1 for analysts assigned to the treatment group and 0 otherwise. *Post* takes the value 1 after the treatment is implemented and 0 otherwise. *Complex* takes the value 1 if the fund is classified as an actively managed equity fund and 0 otherwise. Standard errors are in parentheses, clustered at the analyst level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

A.1 Examples of analyst reports

A.1.1 The summary section of an analyst report that received a high reasoning score

[redacted]'s seasoned management team deftly executes a research-driven, flexible process that delivers income while limiting downside risk. It earns a [redacted] for its cheapest share classes and [redacted] on its more expensive shares. [redacted] brought an income-focused approach to this strategy in late 2011, and after a decade it's clear he has delivered on this strategy's objective. Over his tenure through October 2022, the strategy's 12-month yield averaged 4.8%, impressive compared with the category peer average yield of 2.2%. Though [redacted] takes the lead here, he is well supported. [redacted] joined the manager roster with [redacted], and [redacted] was added in March 2015. The trio's collaboration and ability to tap the firm's extensive resources and strong fundamental teams gives them an edge relative to other multi-asset income peers. Around 17 distinct, well-regarded [redacted] teams run sleeves for the strategy, including the high-yield team that manages [redacted]. Alongside an experienced team is an income approach that differentiates itself through its emphasis on downside protection. Rather than loading up on a mix of risky asset classes that tend to be correlated during stress periods, the team focuses on the correlation of its underlying sleeves to build a well-diversified portfolio. The managers seek to keep the fund's volatility below that of a blended benchmark (50% [redacted] World Index and 50% [redacted] Bond Index) but are not constrained to a strategic allocation. This gives the team a great amount of flexibility, but it has deliberately abstained from making quick and drastic allocation changes that could lead to a whirlwind experience for investors. Instead, the team, with input from the underlying sleeve managers, focuses on its medium- and long-term views to gradually tilt and adjust the portfolio's composition. Management has been able to limit the risk of drawdowns over [redacted]'s tenure. The strategy's downside-capture ratio of 93% relative to the category index compares favorably with the allocation—30% to 50% equity category average of 108%. Investors should be wary, however, of this strategy's credit risk, which may cause bouts of volatility. Around 66% of its fixed-income exposure resides in securities rated BB or below versus around 25% for category peers.

A.1.2 The summary section of an analyst report that received a low reasoning score

[redacted] is a systematic strategy that tilts toward smaller value names, resulting in a portfolio that can exhibit factor sensitivity. With low fees and a strong parent in [redacted], the sole Investor share class retains its [redacted].

The investment team employs a sensible systematic process that seeks to identify stocks that will grow earnings faster than industry peers and that are trading at reasonable valuations. The starting investment universe is its prospectus benchmark, the [redacted] Index. The model looks for stocks that score well on five factors–valuation, growth, quality, momentum, and management decisions–

relative to industry peers. The fund does not make sector bets versus the index but can have certain size and style tilts. The model generally finds more attractive opportunities in smaller value names, relative to the mid-blend category benchmark, the [redacted] Index.

This strategy launched in 1995 as a quantitative fund managed by [redacted], and over time it has adopted a more team-based management approach, similar to other systematic strategies. Currently, the fund is managed by a nine-person alpha equity team, which is part of [redacted]'s 30-person quant equity group. Leadership has traditionally been long-tenured [redacted] professionals, but in 2021, managers [redacted] and [redacted] retired, and [redacted], who joined the group in April 2020, became the sole lead in October 2021. He joined the team with solid quant investment experience, serving as the head of U.S. research at [redacted], where he was responsible for developing, testing, and implementing quant models.

The markets in 2021 have been very favorable to this strategy; in the first 11 months, the fund returned 23.8%, significantly outperforming the category benchmark's 17.8%. This comes after several challenging years when the fund underperformed both its prospectus and the category benchmarks for four consecutive calendar years. In the volatile first quarter of 2020, it underperformed the category index by almost 5 percentage points, owing primarily to an overweighting in small-value names, which as a group declined by about 50%.

A.1.3 Examples of causal statements identified using LLMs

Causal statements	**Cause**	**Effect**
Excluding securitized bonds tilts the fund toward Treasuries and corporates.	Excluding securitized bonds	The fund is tilted toward Treasuries and corporates
The portfolio also tends to have a lower duration than its category index and peers, which makes it less sensitive to changes in interest rates.	The portfolio tends to have a lower duration than its category index and peers	The fund is less sensitive to changes in interest rates
The speed and fluidity at which opinions are translated into prices makes it difficult to consistently beat the market.	The speed and fluidity at which opinions are translated into prices	it is difficult to consistently beat the market
Two comanager departures and a team addition bring some uncertainty to the strategy.	Two comanager departures and a team addition	Uncertainty to the strategy
The strategy's disciplined and conservative credit-driven process has demonstrated its value through time, but the analyst churn casts a shadow on its execution and puts a lid on our confidence level, supporting an Average Process Pillar rating.	The analyst churn	Casts a shadow on the strategy's execution and puts a lid on our confidence level
The China stake rose to 42% from 32% in the quarter ended March 2020, contributing to excellent returns for the remainder of 2020 as China successfully dealt with the virus and outperformed other emerging markets.	The China stake rose to 42% from 32% in the quarter ended March 2020 and China successfully dealt with the virus and outperformed other emerging markets.	This contributed to excellent returns for the remainder of 2020.

A.1.4 Example of an experienced analyst discussing machine-suggested decision driver

The fund shows other, classic signs of bloat: its holdings count has increased and turnover decreased. In 2021, [redacted] noted that a primary reason for the name count to rise by nearly 200 stocks is because, despite it being closed to new investors since 2006, the strategy has swollen. (The other reason he cited was, capital markets have been very active.)

The narrowness of the strategy's outperformance over the past year raises another question surrounding [redacted]'s ability to execute successfully on smaller positions; the strategy outperformed over the past year only because of the stellar gains of its top holding, NVIDIA. Had the stock hypothetically not existed in either the portfolio or Russell 1000 Growth Index, the strategy would have underperformed substantially. (22% vs. 29% gross gain, trailing 1 yr through Nov. 2021.)

But I'm comfortable reaffirming the Above Average [redacted] rating because even though the strategy's size as a share of its investable universe is at an all-time high (see "Strategy Size" chart above), it's apparently not dramatically higher than it was five years ago. Also, [redacted]'s good ideas have been wide-ranging over lengthier periods. Its outperformance over the past five years have come from a variety of overweighted positions – 7, specifically: Amazon.com Inc, Cloudflare Inc, Lululemon Athletica Inc, Moderna Inc, NVIDIA Corp, Shopify Inc, Tesla Inc.

One would need to hypothetically ignore their existence over the past 5 years for the strategy's returns to have matched the index's. That shows that [redacted]'s strong execution is broadly enough based.

We also haven't observed significant profile changes to the portfolio one might detect from a strategy unable to execute as it would like.

A.2 Follow-up experiment and qualitative interviews with financial advisors

We employed a mixed-method approach consisting of an experiment and qualitative interviews to further investigate whether our internal expert ratings of explanation quality generalize across different user populations and contexts. The experiment involves 16 professional financial advisors who represent the firm's most sophisticated institutional investor client base, recruited through our partner company's consumer service vendor. The experiment used a design within subjects in which participants read and evaluated four analyst reports about mutual fund recommendations, with each report varying in explanation quality as previously rated by internal review committees. Participants do not have information on the internal review committee ratings of the analyst explanations. The fund names and identifiers have been masked so that participants cannot search for more information about the funds. The four funds assigned to each participant were randomly selected from a pool of 20 analyst reports on funds that received the same analyst recommendation decision, which is labeled " recommendation with high conviction." Controlling the recommendation decision ensures that participants focus solely on the quality of explanations. The 20 analyst reports in the sample all fall into the US equity funds category to ensure that they are comparable in the investment process. Experiment manipulation involves randomly assigning the presentation order of these four reports to control for sequence effects. Following each report evaluation, participants assess the quality of analyst explanations along multiple dimensions including the usefulness of explanation, clarity of explanation, analyst's demonstrated knowledge of investment strategy, and logical reasoning, before indicating their likelihood to invest in the recommended fund.

After completing the survey experiment, financial advisors also participate in semi-structured 30-minute interviews that explore how they value and use analyst explanations in their professional practice. This design allows us to examine whether our internal expert ratings of explanation quality generalize to external financial advisors, who are important long-term clients of the firm, while the qualitative component provides deeper insights into how professional advisors interpret and apply analyst explanations in real-world decision-making scenarios.

Our interview sample comprised 16 highly experienced financial advisors with substantial tenure and diverse portfolio responsibilities, recruited through a significant investment of resources by both our research team and the collaborating firm. The participants ranged in age from 33 to 70 years, with a median age of 59 years, representing a mature but diverse group of industry veterans. Two of the 16 (12.5%) participants were women. The sample demonstrated exceptional professional experience, with 75% of the respondents having served in their current roles for more than 21 years, indicating deep institutional knowledge within the asset management profession. Portfolio responsibilities varied considerably across participants, encompassing a broad spectrum from boutique operations managing $25-50 million to large institutional portfolios exceeding $1 billion in assets under management (AUM). The median AUM fell within the $100-150 million range, reflecting a balanced representation of asset management operations. Collectively, the participants managed approximately $4.8 billion in total assets, providing substantial market representation and ensuring that

our findings capture the perspectives of professionals responsible for significant investment decisions across diverse portfolio scales and client segments. Given the premium nature of this participant pool, we allocate substantial resources to recruitment and compensation, investing more than $10,000 USD in participant incentives to ensure high-quality engagement from these senior professionals. All interviews were personally moderated by the firm's marketing director and the authors, reflecting the collaborative commitment to data quality and the strategic importance both parties placed on this investigation.

The results of the experiment present strong evidence for the validity of the internal review committees' assessment of analyst explanations. Figure 1 and Table 1 shows that the evaluations by the internal review committees of the analyst explanations are significantly and positively correlated with how the external financial advisors evaluate those explanations along the dimensions of usefulness, clarity, demonstrated knowledge, logic, reasoning and their likelihood of investing in the recommended fund.

The qualitative interviews following the experiment help us to gain a deeper understanding of what financial advisors value in analyst explanations. Consumers of analyst reports seek a structured and analytical approach that provides original insights beyond the compilation of facts. One advisor rated one of the analyst explanations unfavorably because "I found this much more like a book report as opposed to like a book review." As another advisor put it:

> "I want an analyst report that doesn't regurgitate what's already on the rest of your [fund information] page. ... I need to know what's in the background. If somebody were to take this manager out to dinner, spend a weekend on with him in a vacation or something like that, and they started talking business. What would you glean from that? ... Give me the stuff you wouldn't be able to tell from all those numbers."

Despite the importance of acknowledging inherent risks and limitations, advisors prefer a definitive stance that clearly articulates the analyst's recommendation rather than leaving readers to synthesize competing information independently. Several financial advisors particularly criticized certain analyst explanations in their reading samples as lacking clear conviction and hedging opinions.

> "It seemed like the analyst was kind of hedging his opinions a lot. I wasn't quite clear as to what he was exactly saying."
>
> "It's just kind of like hedging. He talks about some great positives, but he also creates some legitimate negatives. So I wish there was more of a stance taken, a little more conviction."

Financial advisors demand analyst opinions to be supported by robust evidence and reasoning that validate their conclusions.

> "There's a lot of opinions in this thing and that's great. I don't mind opinions. I just don't want those opinions to be oxymorons or just filling air."

Financial advisors desire a clear analytical framework that demonstrates how various factors connect and influence the overall assessment, allowing them to understand the logical progression from data points to conclusions.

> "I like more of a logical layout with a a progression of flow and justification for choices made by the manager. ... More information, but condensed in a more concise way and presented in a logical flow."

Specifically, financial advisors want analysts to provide the weighted importance of different factors, explicitly stating which elements matter the most and the reasoning behind their prioritization.

> "Are they all equally weighted? Are they all equally important? I kind of feel like maybe momentum isn't as important if the rest of things are through the roof. Are these (factors) weighted differently? And if so, I'd like to know which is the least important. Say, with values the most important, or qualities most important and low volatility being the least important."

A.3 Pre-registration differences

This study was pre-registered in the AEA Randomized Controlled Trials registry with a pre-analysis plan. All aspects of the experimental sample, design, outcomes, and analyses were implemented without deviation other than the following key differences:

- The pre-analysis plan specified approximately 99 analysts to be part of the experiment, subject to turnover within the company during the experiment. Two analysts left the company before the start of the experiment, and resulted in a sample of 97 analysts.
- The pre-analysis plan describes all possible primary and secondary outcomes, some of which we noted as potentially not being available due to partner and budget constraints. We were indeed not able to obtain some of the outcomes. We also report an additional outcome that was not available pre-experiment and transform one of the pre-registered variables:
  - We pre-specified an outcome measuring reasoning scores additionally provided by MBA students. Due to budget constraints accompanied by the number of reports we would have had to have MBA students score to sufficiently reduce noise, we were not able to undertake this effort. We were also not able to obtain the internal algorithm recommendation quality for each fund rating, although we verified the quality on average.
  - Instead, we obtained an additional variable to measure causal reasoning in the reports, by leveraging Large Language Models to identify causal statements in the draft. We included this result as it provides more insight into how analysts wrote their explanations.
  - We pre-specified an outcome called "complexity of decision drivers", which we natural-log transformed. After we pre-registered the experiment analysis, we learned from the company that reports that clearly articulate its logical path and reasoning process tend to focus on a more selective set of causal drivers instead of having a laundry list of factors. Leveraging this insight from the field, we focus on the number of decision drivers as another way to provide deeper insight into causal explanations.
  - We report all possible exploratory analyses on outcomes in the appendix, as well as heterogeneous treatment effects on gender and prior experience with security. We could not report on two variables that we pre-specified – search for external knowledge and committee request to collect additional information. For the former, analyst responses varied substantially and could not be meaningfully compared quantitatively. For example, analysts' responses ranged from high level answers (e.g."manager interview" and "firm report"), to detailed description of factors and their sources. For the second variable, there were less than 1% of cases where the committee requested to collect additional information, so we did not analyze or report this variable.
- We specified cutting tenure at the median in the pre-analysis plan, but cut tenure by new analysts in the main paper results. We learned after we received the data that three years was the cutoff for new analysts to be promoted officially to analyst levels, as well as the difficulty of learning the company methodology. We report all results cut at the median in the appendix.