12-9-2022

Explainable AI Helps Bridge the AI Skills Gap: Evidence from a Large Bank

Selina Carter
Carnegie Mellon University

Jonathan Hersh
Chapman University, hersh@chapman.edu

Explainable AI Helps Bridge the AI Skills Gap: Evidence from a Large Bank

Comments

This is a preprint of an article that has not yet undergone peer review.

"Explainable AI Helps Bridge the AI Skills Gap: Evidence from a Large Bank"

Selina Carter * Jonathan Hersh†

This draft: December 9, 2022

Abstract

Advances in machine learning have created an "AI skills gap" both across and within firms. As AI becomes embedded in firm processes, it is unknown how this will impact the digital divide between workers with and without AI skills. In this paper we ask whether managers trust AI to predict consequential events, what manager characteristics are associated with increasing trust in AI predictions, and whether explainable AI (XAI) affects users' trust in AI predictions. Partnering with a large bank, we generated AI predictions for whether a loan will be late in its final disbursement. We embedded these predictions into a dashboard, surveying 685 analysts, managers and other workers before and after viewing the tool to determine what factors affect workers' trust in AI predictions. We further randomly assigned some managers and analysts to receive an explainable AI treatment that presents Shapely breakdowns explaining why a model classified their loan as delayed and measures of model performance. We find that i) XAI is associated with greater perceived usefulness but less perceived understanding of the machine learning predictions; ii) Certain AI-reluctant groups – in particular senior managers and those with less familiarity with AI – exhibit more reluctant to trust the AI predictions overall; iii) Greater loan complexity is associated with higher degree of trust in the ML predictions; and iv) Some evidence that AI-reluctant groups respond more strongly to XAI. These results suggest that the design of machine learning models will determine who benefits from advances in ML in the workplace.

1 Introduction

As artificial intelligence continues to replace tasks once solely the domain of humans (Brynjolfsson et al., 2018b), workers are increasingly confronting AI in the workplace. These workplace AI systems vary in their pervasiveness between those that replace individual job tasks to those that reform entire production processes and the roles that are required to operate in them (Frank et al., 2019). The degree to which this is a welcomed change varies both at the individual, company and even sector level. Workers' resistance to AI augmentation can be understood from several perspectives including a status quo bias (Kim and Kankanhalli, 2009), fear of job replacement or loss of power (Acemoglu et al., 2022) but perhaps most perplexingly an overall mistrust of algorithms generally referred to as "algorithm aversion" (Burton et al., 2020). Evidence of

*Department of Statistics and Machine Learning, CMU, shcarter@stat.cmu.edu
†Agyros School of Business and Economics, Chapman University, hersh@chapman.edu

algorithm aversion – and its opposite, algorithmic appreciation (Logg et al., 2019) – has been seen in many domains (Jussupow et al., 2020). Further, understanding the factors that impact algorithm aversion or appreciation will continue to be important for firms and researchers, as advances in machine learning have yet to reach their full potential for their impact on firm level productivity (Brynjolfsson et al., 2018a).

In this paper, we study how workers within a firm respond to a new tool that assists with the presentation of algorithmic decisions, explainable AI (Gunning et al., 2019). Explainable AI (or XAI) is a set of methods applied to an artificial intelligence model or model predictions that help explain why the AI made certain decisions. These can take the form of global explanations, that explain how a model behaves on average, as well as local explanations, that explain how a model made a decision for a particular observation. Partnering with a large infrastructure and development bank, we built a machine learning model to predict when a loan of theirs was likely to be delayed and by how much. For this bank in particular, this is an important problem, as only 22% of the bank's loans were fully disbursed on time. We then built an explainable AI module that presented a delay prediction along with both a performance feedback and local explanation screen that presented Shapely local explanations according to Lundberg and Lee (2017). Next we surveyed 685 employees at the bank, randomizing each employee into either a control group that only received the machine learning predictions, and a treatment group that received the XAI dashboards.

Using a within-subject design, we first ask individuals to report the expected delay for up to two projects in their portfolio. Next we ask them to view the machine learning dashboards, with or without the XAI components, and ask them whether they want to update their delay estimate for their loan. The degree to which they update their loan estimate we view as an indicator of how much they "trust" the algorithm, although we further measure their self-reported understanding and usefulness of the machine learning predictions. Using baseline Poisson regression models to measure how much individuals update their delay estimate with and without the XAI component, we find evidence of algorithmic aversion or lack of updating by "AI reluctant" groups. Specifically, more senior members of the team and individuals with less machine learning experience are much less likely to update their delay estimate after viewing the machine learning predictions. We find in general that XAI results in greater perceived usefulness and less perceived understanding of the machine learning predictions although this is also moderated by user characteristics. We find that greater loan complexity, measured by total loan size, is positively associated with trust in machine learning predictions. We also find evidence that these AI-reluctant groups respond more strongly to XAI. Namely, if individuals with more seniority and less machine learning familiarity are in the XAI treatment, they are much more likely to update their delay estimate, that is trust the AI predictions.

2 Background and Hypotheses

2.1 Related Literature

This papers speaks to three broad areas of research. The first is the recent but growing literature on factors that drive algorithmic trust including trust in decisions made fully or partially by artificial intelligence. This literature has examined several factors that may increase or decrease algorithmic trust including overall performance of the model, transparency of the algorithm used, the degree to which the model is interpretable, whether a user has some degree of control over the algorithm, and several user-specific characteristics such as user age, or familiarity with machine learning (Jussupow et al., 2020). Dietvorst et al. (2015) suggest that users distrust algorithms when seeing them act in error and Dietvorst et al. (2018) suggests that trust is impacted by the degree to which users can modify them. Even the display of the avatar of the algorithm has been associated with trust. Ganbold et al. (2022) find that more competent looking avatars reduce algorithmic aversion in the context of financial advice.

The second literature that this paper contributes to is the growing body of work on how management styles and management processes may adjust within the firm in the age of algorithms. By replacing some tasks previously performed by the worker, the optimal worker-manager relationship may shift from one of top-down control, to one characterized by lateral feedback and coaching. Additionally, workers may resent or prefer being managed through algorithms, which they may view as either more or less fair depending upon their use case and deployment. Tong et al. (2021) implement a field experiment in a financial services firm, where performance feedback for employees is delivered and generated either by an AI or a human manager. The authors find that the AI system provided better feedback on employee performance, and resulted in a higher job performance rating compared to human ratings. However, upon hearing that the ratings were generated by an AI system, some workers responded negatively, slightly lowering job performance ratings. Finally our work speaks to research on the changing human capital demands brought by AI. Specifically, research are beginning to turn their focus towards how firms will retrain workers whose jobs will require more AI competence (Benzell et al., 2022).

2.2 Hypotheses

Distrust or aversion to AI could have many sources, and entangling these underlying mechanisms will require carefully designed experimental conditions and hypotheses.

Informed by previous research we investigate five hypotheses to determine how explainable AI will be used by managers within firms. The first of these hypothesis is:

H1: Workers who receive the explainable AI treatment will exhibit more trust in the AI overall.

While hypothesis **H1** estimates the effect on average across all participants, we consider that some workers may exhibit more baseline trust in AI than others. This may be due to several factors. Consider a worker that possesses relatively high complementary skills to AI, for example workers who have statistical or programming knowledge. These workers may view AI as labor augmenting of their particular tasks performed in their roles (Goldin et al., 2020). A worker who is more familiar with AI in general – through either having studied or used it in the past – is also more likely to understand how it functions. Familiarity itself has been long documented to be positively related to consumers' trust in any new technology (Komiak and Benbasat, 2006). Mapping these theories onto our data, in our survey we captured the role that a worker performs for the loan, which vary from the technical to the non-technical. We also capture their self-reported familiarity with AI/ML on a Likert scale. This leads us to the formulation of our second and third hypotheses:

H2: Technical experience will have a moderating effect on AI trust. More technical roles, such as analysts, will exhibit more trust in the algorithm, as will individuals who self report more AI/ML familiarity.

H3: More senior members of the team will be less likely to trust the AI overall.

While the heterogeneity of the workers using AI might affect the outcome, we consider that the heterogeneity of the tasks to which we are applying AI might be an important moderator. Consider the puzzle that users seem to exhibit varying preferences over AI automation of decision making (Fuchs et al., 2016). On one hand, users of a technology sometimes exhibit too much trust and reliance on AI decision-augmenting tools, a so called "automation bias." This has been documented in medical decision-making, such as the case of automated electrocardiogram diagnostic tools, where users do not disagree with the tool even if it is obviously in error (Bond et al., 2018). On the other hand, technology users exhibit low trust for some automated systems, such as for vehicle piloting systems (Stocker, 2022), or for robotic surgery (Sullins, 2014).

There is limited theory to suggest what the important heterogeneity in our context that might affect trust. While Venkatesh et al. (2012) suggest that age, gender, and experience are important moderators for general technology acceptance, it is unknown if these translate to the context of algorithmic trust. We hypothesized that an important feature for the loan is the overall size of the loan. A very large loan indicates a higher degree of complexity. This may lead a manager to trust the AI suggestions more, given that the AI may suffer less from cognitive overload and be able to consider a loan with many features. However, large loans also may indicate a higher consequence of mis-classification. An analogue here might be to the self-driving car, where the consequence the AI self-driving car's actions can be extremely costly to the user, which may be a factor affecting low trust in automated self-driving cars. On net we believe the latter effect dominates leading to our fourth hypothesis:

H4: Managers will be less likely to trust the AI's decisions for more important loans.

Finally we consider who stands to benefit more from explainable AI – those more familiar with AI or those with less? The manager-level moderators of **H2** and **H3** suggest there are "AI-reluctant" groups who display less baseline trust of AI. This could be due to lacking AI-complementary skills, or due to concern of task or job replacement[1]. On the latter point, Brynjolfsson et al. (2018b) find that the degree of AI task replacement varies quite widely across jobs. If automated AI systems stand to replace tasks, these workers may display reluctance to trust these systems. However, if they are more concerned with lacking complementary skills then providing explainable AI modules provides AI model output in a form that is more easily digestible for these groups. This leads us to our final hypothesis:

H5: AI reluctant groups will relatively display more trust for AI when accompanied by the explainable AI module.

3 Empirical Setting and Machine Learning Model

3.1 Empirical Setting

We partnered with a major development bank to build a machine learning model that could predict delays in execution of sovereign guaranteed investment loans. The bank in question offers these loans to governments in the Americas and are typically used for the purpose of large infrastructure projects that take many years to build. Often the original timeline for these loans is not followed, for reasons internal and external to the bank. The disbursement of loan funds (i.e., the principal amount) is incremental and depends on achieving key project milestones. Loan disbursements are usually scheduled to finish within five years. However, only 22% of loans disburse all funds on or before their scheduled time, meaning 78% of loans in the bank's historical portfolio experienced some form of delay. These delays can be quite costly. The average loan that is delayed is setback by 14 months. The bank estimates that 24% of their total supervision costs are from periods when the loan is technically delayed, that is, past its "original disbursement deadline." These delays can be somewhat surprising as well, as managers sometimes report their projects are on track until suddenly encountering difficulties just before the disbursement deadline. Clearly the bank could reduce its overall costs if it had more accurate methods to identify risky projects, particularly during the design phase.

Further complicating the matter for this firm, the average size of these loans is quite large.

[1]One concern with moderators is that the search to identify significant effects occurs ex-post rather than ex-ante, resulting in incorrect confidence intervals. For all of the moderators discussed here we publicly pre-registered the analyses available at https://aspredicted.org/um8s7.pdf.

The average size of the loan is $67m USD in 2020 dollars. The bank has around 90 new loans per year, and roughly 500 active loans on its balance sheet during the year.

3.2 Machine Learning Model to Predict Loan Delays

Working with the bank, we developed a model to ingest characteristics on the loan, the country where the loan was being deployed, and on the team handling the loan, and predict the estimated months the loan would eventually be delayed[2]. They desired a model that was dynamic and could produce updated delay estimates every month as new information on the development of the loan was provided. The model we developed was of the form

$$y_{iT} = f(X_{i_t}|\Theta_t) + \epsilon_{it}$$

where y_{iT} is the number of months the loan is expected to be delayed at final disbursement time T. The matrix X_{it} holds all predictor characteristics about the loan, country and team at the bank handling the loan known at time t. We consider Θ_t, the information set, to be all information about the loan and the country known at time t. Finally ϵ_{it} is the error. Note that the statistical model has a forecasting quality, whereby the goal is to determine, based on the information set Θ_t, the stochastic outcome at the eventual end date T. For more details on the machine learning model, a full list of input features, and performance metrics, we refer the interested reader to the Appendix section B.

4 Experimental Design

We presented the machine learning model to the bank in early 2021, who allowed us to embed the predictions into a series of dashboards with the intention to randomize an explainability module that was presented along with the predictions. Individual responses were collected via an online survey, which we sent to 685 participants on June 14 of 2021. The experiment concluded on July 21st of 2021. A full list of the questions asked in survey is presented in the Appendix section C. Individuals were first asked a series of questions about their background, including experience with machine learning, job role, etc. We then asked them to list up to two projects to which they were directly assigned. Although their familiarity with each project may vary, the number of loans a team member is assigned is usually in the single digits. Thus they should be familiar enough with projects listed to give pertinent details as to the likelihood of delay even in absence of a machine learning suggestion.

For each of the projects listed, we asked respondents to indicate their best estimate as to

[2] A full list of features used in the machine learning model are presented in B.3

the number of months the loan is likely to be delayed. From here, we presented participants with either the control or treatment machine learning prediction dashboards, along with a 4-10 minute explanation video, in both English and Spanish, as to how to read and interpret the ML prediction dashboard. After asking the participant to locate their project and view the expected machine learning delay, we then asked the respondent if they would like to update their delay estimate, and if so by how much.

4.1 Explainability Randomization

Figure 1 presents the dashboard viewed by the control group. These individuals received information about predicted delays for their projects, along with prediction confidence intervals that show the degree of certainty in the model's prediction. Treated groups received this dashboard along with two additional screens, shown in figures 2 and 3. These screens presented both information about the overall model accuracy as well as the local interpretability of the model. The model accuracy screen presented scatter plots of predicted and true delay estimates for historical loans. The explainability screen presented Shapley local explanation a la Lundberg and Lee (2017). We collectively refer to screens in the treated group dashboards in figures 2 and 3 as the "explainability" treatment, keeping in mind that they measure both local explainability and model performance[3].

4.2 Survey Summary Statistics

Table 1 presents summary means, min, max and standard deviations over the resulting survey data. We submitted the survey to 685 participants and received 490 responses. We removed any individuals who could not find any projects under their supervision in the machine learning dashboard, leaving us with 453 distinct individuals. Of these, 174 listed one project under their supervision, and 279 listed two projects under their care. This generates the full experimental sample of 732 project-users shown in table 1.

The summary statistics shows other characteristics captured during the survey. We collected the loan value, which has an average value of 17.84 log US dollars. Most users in the survey were somewhat familiar with machine learning, with only 11% reporting no familiarity with ML. Conversely 5% reported being very or extremely familiar with machine learning. We additionally captured the role the responded performed for the loan. A team covering a loan can be comprised of several analysts and procurement and fiduciary specialists. The team leader and alternate team leader oversee the loan operations from the bank side. Finally, the chief of operations is the

[3]While we would have preferred to have the explainability and model performance dashboards to be in separate treatment arms, early power calculations showed that we lacked the requisite statistical power to identify an effect with 80% power using three treatment arms. The authors then decided to combine the two treatment arms to ensure 80% power coverage.

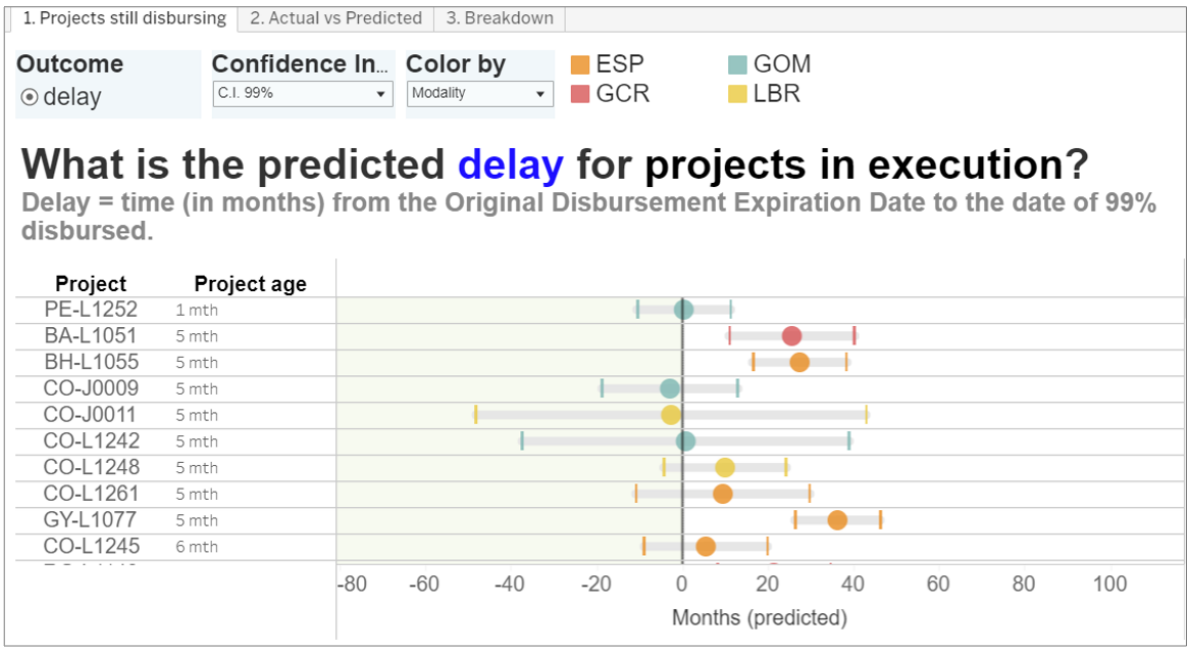


Figure 1: Prediction screens for the control group. Control participants received only the dashboard screen shown here. This screen shows the predicted estimated months of delay in execution for the project, along with the confidence intervals for that prediction. Confidence intervals were generated via Infinitesimal Jackknife (Wager et al., 2014)

highest leadership role and is responsible for most client-facing interactions. We see that 31% of projects are covered by analysts, 12% by fiduciary or procurement. In terms of managers, 39% occupy a team leader position and 4% of them are chief of operations. We also survey individuals on the degree to which they understood the machine learning model on a Likert scale from one to five. Only 1% of respondents report not understanding the tool at all, and 2% report understanding it only "a little." The modal response, at 43%, is that respondents understood the tool "easily."

The main outcome variable we will capture is the absolute value of the change in delay estimate before and after viewing the machine learning tool. The summary statistics table shows that on average, respondents changed their delay estimate by 2.61 months, with the modal response being 0 months of delay estimate. We interpret that individuals who change their delay estimate by larger amounts show more "trust" in the machine learning model.

As some model-free evidence of **hypothesis 1** we present a scatter plot of the respondents' delay estimate before and after viewing the machine learning dashboards in figure 4. On the x-axis we plot the delay estimate before viewing the tool, and on the y-axis we show their delay estimate after viewing the tool. A user who exhibits no trust or belief in the machine learning predictions will have their delay estimate lie perfectly on the 45 degree line, that is their guess before and after viewing the tool is the same. Deviations away from the 45 degree line indicate

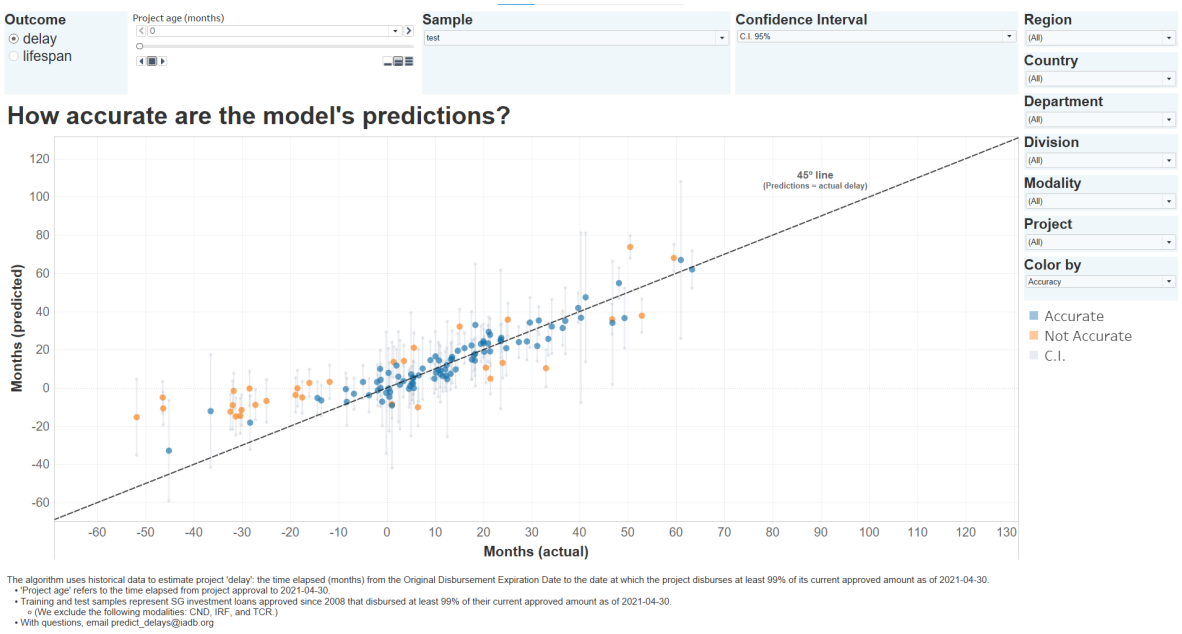


Figure 2: Additional screens for explainable AI treatment group. Model performance panel. Model predicted months of delay are presented on the y-axis along with the actual months delayed on the x-axis. Note these plots only present historical projects and not current projects.

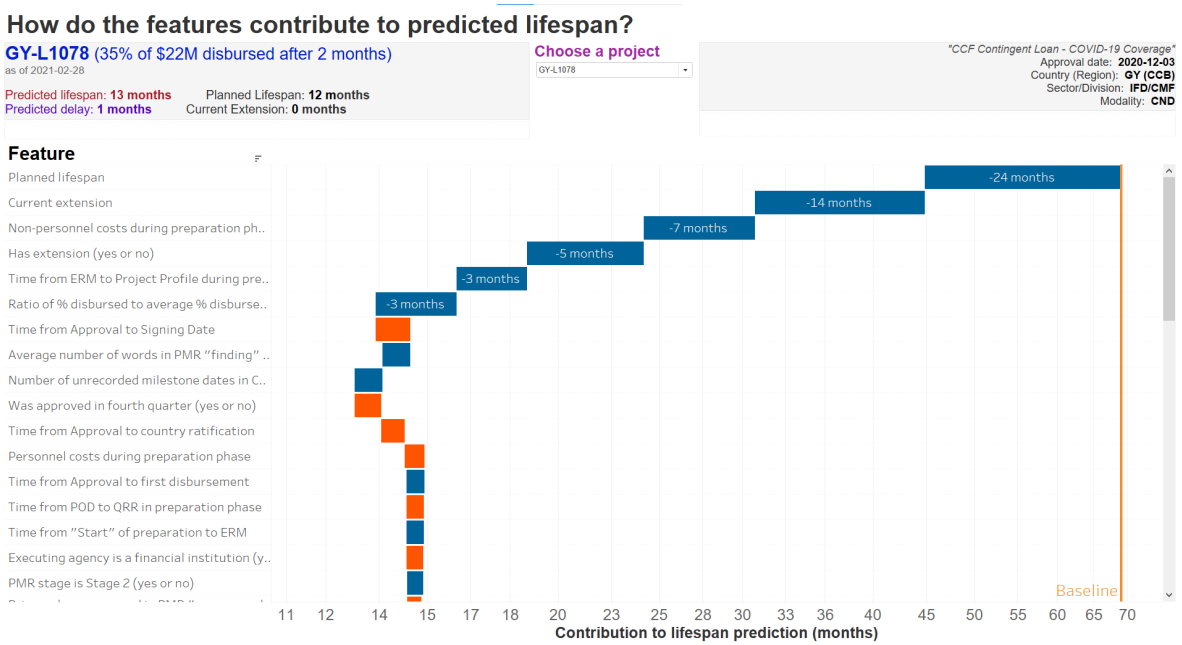


Figure 3: Additional screens for explainable AI treatment group. Model explanation panel. Shapley Additive explanation (SHAP) according to Lundberg and Lee (2017) are presented, showing how the model predictions change when conditioning on the given feature. SHAP values are presented in a common "waterfall" presentation, ordered from largest to smallest change in expected value.

a greater willingness to trust the machine learning predictions, as individuals have been induced to update their delay estimate by the predictions.

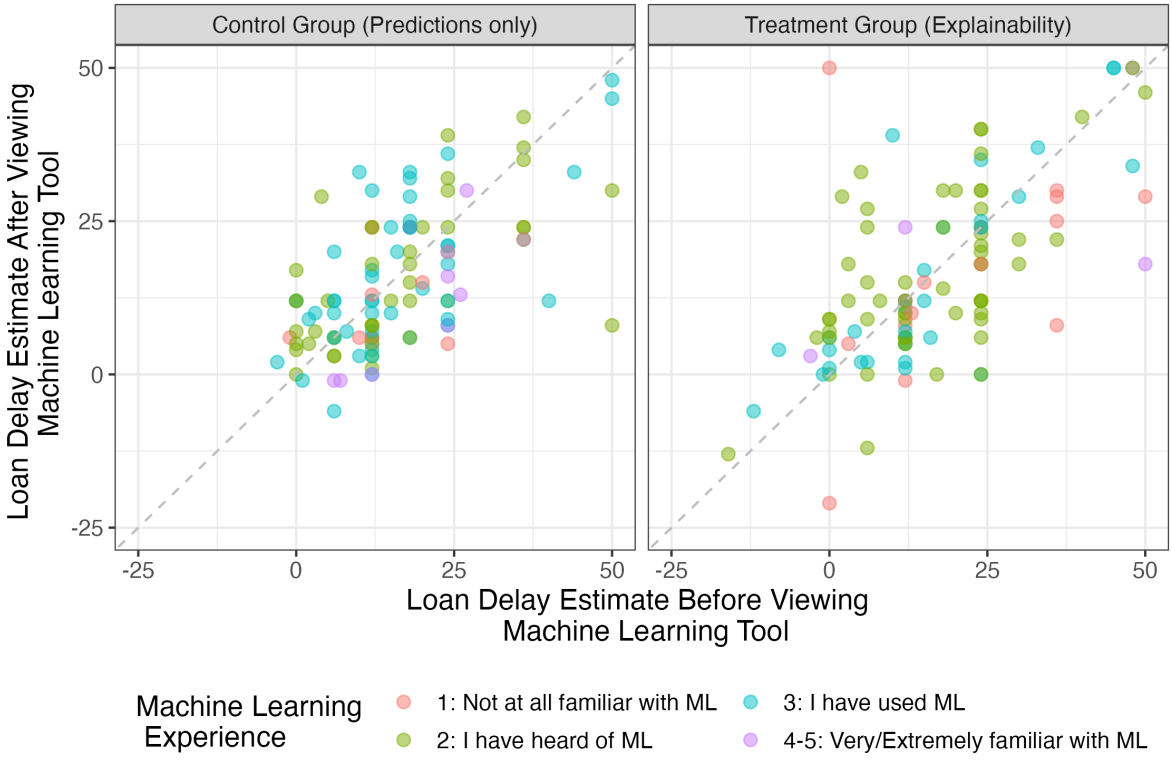


Figure 4: Loan Delay Estimate Before/After Viewing Machine Learning Tool

The figure plots out the distributions for the control and treatment groups. We see in the right panel, for the explainability treatment group, a greater dispersion around the 45 degree line than seen on the left panel. This indicates that individuals who receive the explainability treatment may be more willing to trust the AI and update their delay estimate. We also see some model-free evidence for **hypothesis 5**, that AI reluctant groups, specifically those without machine learning familiarity, display more trust in ML when the predictions are accompanied by explainability. The same figure colors each point by the respondent's self-professed degree of machine learning familiarity. For the prediction only group in the left panel, we see individuals with little familiarity in machine learning tend to do very little updating to their delay estimate after viewing the ML predictions. However, for the panel on the right, the explainability sample, we see individuals without ML familiarity updating their beliefs by much larger amounts. Of course to test this formally will require more rigorous statistical tests, which we will turn to in the next section.

5 Empirical Results

To determine the impact of the explainable AI we begin with a formal series of statistical tests. Table 2 presents our baseline estimate. Because the outcome variable – the number of months in absolute value a user updates their delay estimate after viewing the ML dashboards – is a discrete count variable, we use as a baseline method a Poisson regression model. The baseline equation we run is of the form $E[y_i|X_i] = e^{\beta^T X_i}$, where y_i measures an individual i's change in delay estimate after viewing the machine learning tool, X_i is a vector of moderators at the individual and loan level that may influence the amount by which they update, and β are the coefficients for those characteristics.

Column one in table 2 regresses an indicator variable for whether the individual is randomized into the explainable AI treatment on the amount a user updates their delay estimate. We find a significant coefficient of 0.129. The magnitude suggest that an individual updates their estimate by an additional 13% if they happen to be in the explainable AI treatment. Given that the average user updates their estimate by 2.61 after viewing the tool, this represents a small but meaningful impact on a measure of trust in the machine learning algorithm. The second column adds to the treatment indicator dummy indicators pertaining to the role the user performs on the loan team. We see that more more senior members of the team – specifically the chief of operations, specialists, and team leaders – are less likely to update their delay estimates when viewing the machine learning predictions dashboards. The most senior member of the team, the person acting as the client-facing chief of operations, updates their estimate by 70% less than an analyst, which is the base level of the factor in the regression equation. This represents a substantial decrease in the trust placed in the machine learning model. We see the coefficient on the XAI treatment declines to 0.089 but remains statistically significant. Column three in table 2 adds the log of the total amount of the loan to the explainable AI treatment indicator. The coefficient on the amount of the loan is 0.146 indicating employees covering larger loans are more likely to trust the ML predictions, all else being equal. We see the XAI coefficient in this regression remains significant in similar magnitude of column 1. Column four adds to the XAI treatment indicator dummy variables that measure self-reported familiarity with machine learning. We see negative and significant coefficient on the indicator whether someone has used ML, indicating moderate levels of machine learning familiarity are associated with less trust in ML.

Overall, the results in table 2 suggest evidence for several of the hypotheses we outlined earlier. First, we see evidence that XAI engenders more trust in AI, providing evidence in favor of **hypothesis 1**. We also find evidence that more senior members of the team display less trust in AI overall, confirming **hypothesis 2**. We also find some evidence in favor of AI being trusted

for more complex environments such as larger loan, evidence in favor of **hypothesis 4**. We have somewhat weaker evidence for the role that technical experience, specifically machine learning familiarity, plays in trust in machine learning systems, that is weaker evidence for **hypothesis 2**. We may have additional concerns with statistical robustness, which will be explored in table 6.

5.1 Impact of XAI on Perceived Usefulness and Model Comprehension

We may also be interested in the impact of XAI on other outcomes, such as how useful the model is perceived to be, and the degree to which XAI assists in the understanding of the model. Figure A1 plots the ordered outcome variable, where we elicit from users their Likert scale for how useful they perceived the machine learning model to be. Most users found the machine learning tool either somewhat or very useful, and this seemed to vary with machine learning familiarity, with those at the ends of the distribution of machine learning familiarity being more likely to find the tool helpful.

Table 3 explores this concept using a bit more rigor. The table shows impact of XAI on perceptions of usefulness, regressing the moderators in the previous table on the outcome variable of perceptions of usefulness of the model. Because the outcome variable is an ordered factor variable, we use an ordered logistic model to estimate the results. When regressing the XAI dummy on the ordered usefulness variable alone, we find the coefficient on XAI to be positive 0.29 and significant. This indicates that individuals in the XAI treatment perceived explainable AI to be more useful. When including role dummies, the XAI treatment indicator becomes marginally insignificant along with most of the role dummies, although in the full specification in column 5 the XAI treatment is again significant. The team leader coefficient remains strongly negative and significant. Log of total amount of the loan does not appear to have an impact on the perceptions of usefulness. When we regress all the moderators together we find a significant coefficient on the XAI treatment, the roles of chief of operations and team leader to be negative and significant, and those most familiar with ML coefficient to be significant and positive.

We also might be interested in the degree to which XAI contributes to perceptions of model understanding. We asked users to report on a Likert scale how well they understood the machine learning dashboards, the results of which are plotted in figure A2. Most users report understanding the dashboards either "easily" or "most." What's interesting here is that the XAI group had fewer users reporting understanding the model at the highest level of "easily." Machine learning experience seems to be correlated with perceptions of understanding, as those with the highest level of machine learning experience most likely to select the highest levels of understanding.

Table 3 test this more rigorously, regressing the moderators and XAI treatment indicator against the Likert scale of understanding. Across the specifications, the XAI treatment has a negative impact on perceived understanding. The coefficient varies between −0.838 and −0.866 indicating a strong and negative impact. Why users find models with XAI to be harder to understand is a little puzzling. Presumably more explanation for why the AI is making certain decisions will lead to a greater understanding of the model predictions themselves. On the other hand, perhaps the additional screens increased cognitive overload and users found it difficult to absorb all of the information presented. Across the specifications there appears to be no impact of worker role or size of the loan on understanding of the machine learning model. There appears to be a strong relationship between machine learning background and perceptions of understanding, with the more ML familiar users reporting much more understanding of the model in general.

5.2 Heterogeneous Treatment Effects of XAI: Which Workers Benefit the Most?

We next to the questions outlined in **hypothesis 5**. Table 2 shows evidence that certain groups are reluctant to trust machine learning predictions in general. We find these characteristics at the role level – specifically more senior roles such as chief of operations or team leads – and from experience or educational backgrounds – those who have less experience with machine learning. To explore how these reluctant groups specifically respond to XAI we return to the regression equation in table 2. We next interact all of the moderator variables with an indicator for whether the observation is in the XAI treatment, resulting in a regression equation of the form $E[y_i|\tilde{X}_i, D_j] = e^{\tilde{\beta}^T * 1_{(D_j=1)} * \tilde{X}_i}$, where D_j is an indicator for whether a user is part of some moderator group j. This regression equation will generate heterogeneous treatment effects for all of the moderators, indicating the difference in impact of each moderator between the XAI and non-XAI treatment groups.

Table 5 shows the results. Exploring how each role responses differently to the treatments, we see that the largest coefficient on role is for the chief of operations. This indicates that the chief of operations role is the most likely to change their update on the ML model's suggestion if there is an XAI component included. Similarly, we see positive and significant coefficient on the lowest level of machine learning familiarity, those that indicate they have no familiarity with machine learning or have only heard of ML[4].

[4]Note we have changed the base level of this factor to produce a coefficient for the lowest level of machine learning familiarity. The base level of the factor is now the highest level of machine learning familiarity.

5.3 Robustness to Alternative Specifications

While we have found evidence that XAI affects algorithmic trust, we next consider how different specifications may alter our understanding of the impact of XAI on trust. Table 6 presents several alternative specifications of the baseline model. The first specification includes a fixed effect at the loan level. While the assignment of the XAI treatment is random, perhaps there are characteristics at the loan level that are correlated with the individual response. We see the coefficient on the XAI treatment increases to 0.207 and remains statistically significant at the traditional levels. The second columns adds all the user-specific controls along with the loan fixed effects. We find the coefficient is marginally insignificant in this specification but the magnitude is comparable to the specification without the user fixed effects. The third column adds an indicator for whether a user has two or more projects in their portfolio and includes loan fixed effects. Recall that in our survey, 174 individuals had one project under their supervision, and 279 listed two projects. Under this specification the coefficient is estimated at 0.212 and remains significant.

The fourth column adds two variables intended to capture the "surprise" of the ML prediction, as well as a measure for whether the user is well-calibrated to the empirical likelihood of delay. The first additional parameter measures the difference between the ML prediction of delay and the planned loan lifespan. We would imagine that if the machine learning prediction deviates significantly from the planned lifespan this may be more novel information for the user, and might affect their degree of trust. We find a small and insignificant coefficient, indicating that users are unlikely to find information of this sort to have a large affect on algorithmic trust. The second added variable adds the surveyed information on what the user estimates as a typical delay for a loan. The intention here is that perhaps users who are poorly calibrated – that is they are unaware as to the empirical delays experienced by the loans in the typical portfolio – would be more or less likely to trust the algorithm. In general we found workers on average to be well calibrated to the empirical likelihood of delays. But how does bias perform on an individual level? We find that the bias variable has a small and insignificant coefficient, indicating those who are poorly calibrated to the empirical likelihood of delays fare no better with the ML predictions than those who are well-calibrated, at least in this context.

The last column regresses the XAI treatment on the amount of delay updating and includes user fixed effects. We see the standard errors grow large and the estimated coefficient is 0.117, not far from the baseline estimates in table 2. Because the data includes individuals with both single and multiple projects, we cannot property identify user fixed effects, although our experimental design differences out large deviations in user delay estimates.

6 Conclusion

Algorithmic assisted decision making is still in its infancy. The rise of large language models such as (Floridi and Chiriatti, 2020) and of diffusion-based image generation (Song et al., 2020) suggest that there are many processes within businesses that will be transformed in the coming decades. Many of these processes will lend themselves to advisory algorithms, that will involve decision-making by a human aided with an algorithm. Blind reliance on the algorithm may lead to less preferable outcomes (Luo et al., 2019), and in the best outcome is likely achieved through a blend of human and algorithmic interaction (Gunaratne et al., 2018)

This paper finds evidence that explainable AI may be an important tool in the design of algorithmic advisory systems. We used a field experiment to study how a large bank uses machine learning predictions with and without a randomized explainable AI treatment. We find evidence that XAI matters: XAI aids in algorithmic trust as measured by users altering their behavior based on the AI predictions. Users also report XAI to be useful although paradoxically decreases their self-reported understanding of the AI. XAI also matters more for different groups. While unconditionally, those with more machine learning experience are more likely to trust AI, if users are randomized into the XAI treatment, they are much more likely to update their beliefs based on the AI predictions.

The results in this paper add an important data point to the concern about algorithmic avoidance. Namely that that choice for how to present the algorithmic predictions plays an important role as to how much they are trusted in a human and advisory algorithm scenario. The choice of the design of the systems also has implications for who within a firm can benefit from artificial intelligence. The default choice without XAI may render advances in AI only for those with considerable AI experience. Our results suggest that gains from AI can be broad, provided we have sufficient transparency within the algorithms and allow workers to combine their own knowledge with the intelligence of the algorithms.

References

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo, "Artificial intelligence and jobs: evidence from online vacancies," *Journal of Labor Economics*, 2022, *40* (S1), S293–S340.

Benzell, Seth G, Erik Brynjolfsson, and Guillaume Saint-Jacques, "Digital Abundance Meets Scarce Architects: Implications for Wages, Interest Rates, and Growth," 2022.

Bond, Raymond R, Tomas Novotny, Irena Andrsova, Lumir Koc, Martina Sisakova, Dewar Finlay, Daniel Guldenring, James McLaughlin, Aaron Peace, Victoria McGilligan et al., "Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms," *Journal of electrocardiology*, 2018, *51* (6), S6–S11.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics," in "The economics of artificial intelligence: An agenda," University of Chicago Press, 2018, pp. 23–57.

_ , Tom Mitchell, and Daniel Rock, "What can machines learn, and what does it mean for occupations and the economy?," in "AEA papers and proceedings," Vol. 108 2018, pp. 43–47.

Burton, Jason W, Mari-Klara Stein, and Tina Blegind Jensen, "A systematic review of algorithm aversion in augmented decision making," *Journal of Behavioral Decision Making*, 2020, *33* (2), 220–239.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114.

_ , _ , and _ , "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, 2018, *64* (3), 1155–1170.

Floridi, Luciano and Massimo Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines*, 2020, *30* (4), 681–694.

Frank, Morgan R, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro et al., "Toward understanding the impact of artificial intelligence on labor," *Proceedings of the National Academy of Sciences*, 2019, *116* (14), 6531–6539.

Fuchs, Christoph, Christian Matt, Thomas Hess, and Christian Hoerndlein, "Human vs. Algorithmic recommendations in big data and the role of ambiguity," 2016.

Ganbold, Odkhishig, Anna M Rose, Jacob M Rose, and Kristian Rotaru, "Increasing Reliance on Financial Advice with Avatars: The Effects of Competence and Complexity on Algorithm Aversion," *Journal of Information Systems*, 2022, *36* (1), 7–17.

Goldin, Claudia, Lawrence F Katz et al., "Extending the race between education and technology," in "AEA Papers and Proceedings," Vol. 110 2020, pp. 347–51.

Gunaratne, Junius, Lior Zalmanson, and Oded Nov, "The persuasive power of algorithmic and crowdsourced advice," *Journal of Management Information Systems*, 2018, *35* (4), 1092–1120.

Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang, "XAI—Explainable artificial intelligence," *Science robotics*, 2019, *4* (37), eaay7120.

Jussupow, Ekaterina, Izak Benbasat, and Armin Heinzl, "Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion," 2020.

Kim, Hee-Woong and Atreyi Kankanhalli, "Investigating user resistance to information systems implementation: A status quo bias perspective," *MIS quarterly*, 2009, pp. 567–582.

Komiak, Sherrie YX and Izak Benbasat, "The effects of personalization and familiarity on trust and adoption of recommendation agents," *MIS quarterly*, 2006, pp. 941–960.

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

Lundberg, Scott M and Su-In Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 2017, *30*.

Luo, Xueming, Siliang Tong, Zheng Fang, and Zhe Qu, "Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases," *Marketing Science*, 2019, *38* (6), 937–947.

Song, Jiaming, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

Stocker, Alexander, "Driver Trust in Automated Driving Systems," *ECIS 2022 RESEARCH PAPERS*, 2022.

Sullins, John P, "Ethical trust in the context of robot assisted surgery," in "AISB 2014-50th Annual Convention of the AISB" 2014.

Tong, Siliang, Nan Jia, Xueming Luo, and Zheng Fang, "The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance," *Strategic Management Journal*, 2021, *42* (9), 1600–1631.

Venkatesh, Viswanath, James YL Thong, and Xin Xu, "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology," *MIS quarterly*, 2012, pp. 157–178.

Wager, Stefan, Trevor Hastie, and Bradley Efron, "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife," *The Journal of Machine Learning Research*, 2014, *15* (1), 1625–1651.

Tables

Table 1: Summary Statistics

Variable	N	Mean	Std.Dev	Min	Max
Treatment group					
... Control Group (Predictions only)	732	0.54	0.5	0	1
... Treatment Group (Explainability)	732	0.46	0.5	0	1
Log of total loan value ($USD)	732	17.84	33.16	15.61	20.21
Machine learning familiarity					
... 1: Not at all familiar with ML	732	0.11	0.31	0	1
... 2: I have heard of ML	732	0.45	0.5	0	1
... 3: I have used ML	732	0.39	0.49	0	1
... 4-5: Very/Extremely familiar with ML	732	0.05	0.23	0	1
Worker role					
... Other	732	0.13	0.34	0	1
... Chief of Operations	732	0.04	0.2	0	1
... Fiduciary/Procurement	732	0.12	0.32	0	1
... Operational Analyst	732	0.31	0.46	0	1
... Team Leader	732	0.39	0.49	0	1
Worker location					
... HQ/Other	732	0.15	0.36	0	1
... Country Office	732	0.85	0.36	0	1
Change before/after using ML tool in absolute value	732	2.61	183.24	0	53
How well did you understand the machine learning tool?					
... 1: Not at all	732	0.01	0.08	0	1
... 2: A little	732	0.02	0.12	0	1
... 3: More or less	732	0.16	0.37	0	1
... 4: Most	732	0.39	0.49	0	1
... 5: Easily	732	0.43	0.5	0	1
How useful is the machine learning tool?					
... 1: Not at all useful	732	0.03	0.17	0	1
... 2: Not very useful	732	0.13	0.33	0	1
... 3: No opinion	732	0.1	0.29	0	1
... 4: Somewhat useful	732	0.51	0.5	0	1
... 5: Very useful	732	0.24	0.43	0	1

Notes: Summary statistics for survey responses shown.

Table 2: Impact of Explainable AI on ML Trust/Delay Estimate Updating

	(1)	(2)	(3)	(4)	(5)
Explainable AI treatment	0.129***	0.089*	0.121***	0.105**	0.052
	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)
Role: Chief of Operations		−1.232***			−1.218***
		(0.186)			(0.187)
Role: Fiduciary/Procurement		0.017			0.010
		(0.070)			(0.071)
Role: Other		0.189***			0.171**
		(0.071)			(0.072)
Role: Specialist		−0.578***			−0.591***
		(0.150)			(0.151)
Role: Team Leader		−0.526***			−0.531***
		(0.057)			(0.057)
Log Total Amount of Loan ($USD)			0.146***		0.161***
			(0.025)		(0.025)
ML Familiarity 2: Have heard of ML				−0.087	0.011
				(0.072)	(0.073)
ML Familiarity 3: Have used ML				−0.404***	−0.321***
				(0.076)	(0.078)
ML Familiarity 4-5: Very/Extremely familiar with ML				0.078	0.149
				(0.107)	(0.109)
Constant	0.899***	1.125***	−1.715***	1.089***	−1.636***
	(0.032)	(0.044)	(0.443)	(0.068)	(0.445)
Observations	732	732	732	732	732
Log Likelihood	−3,352.430	−3,255.340	−3,334.816	−3,324.693	−3,210.737
Akaike Inf. Crit.	6,708.860	6,524.681	6,675.633	6,659.386	6,443.473

Notes: Poisson model estimates shown. Dependent variable is the absolute value of the change in loan delay estimate after viewing ML predictions. Larger values indicate a greater willingness to trust the ML model's predictions relative to the employees first estimate, assessed prior to using the ML tool. Explainable AI treatment refers to the treatment effect of being randomly assigned into the group that received the additional screens that gave information on model performance and model explanations. *p<0.1; **p<0.05; ***p<0.01

Table 3: Impact of Explainable AI on Perceived Machine Learning Usefulness

	(1)	(2)	(3)	(4)	(5)
Explainable AI treatment	0.290**	0.230	0.295**	0.287**	0.234*
	(0.140)	(0.141)	(0.140)	(0.141)	(0.142)
Role: Chief of Operations		−0.526			−0.582*
		(0.336)			(0.339)
Role: Fiduciary/Procurement		−0.298			−0.331
		(0.241)			(0.244)
Role: Other		−0.081			−0.076
		(0.259)			(0.260)
Role: Specialist		−0.105			−0.122
		(0.407)			(0.406)
Role: Team Leader		−0.931***			−0.957***
		(0.171)			(0.172)
Log Total Amount of Loan ($USD)			−0.078		−0.084
			(0.074)		(0.076)
ML Familiarity 2: Have heard of ML				0.117	0.251
				(0.239)	(0.242)
ML Familiarity 3: Have used ML				0.172	0.346
				(0.241)	(0.246)
ML Familiarity 4-5: Very/Extremely familiar with ML				0.471	0.704*
				(0.383)	(0.389)
Observations	732	732	732	732	732

Notes: Ordered logistic model coefficients shown. Dependent variable is the Likert scale response to the question: "How useful is the machine learning tool?". Larger values indicate greater perceived usefulness in the ML tool. Explainable AI treatment refers to the treatment effect of being randomly assigned into the group that received the additional screens that gave information on model performance and model explanations. *p<0.1; **p<0.05; ***p<0.01

Table 4: Impact of Explainable AI on Perceived Understanding of Machine Learning Model Predictions

	(1)	(2)	(3)	(4)	(5)
Explainable AI treatment	−0.838***	−0.834***	−0.838***	−0.862***	−0.866***
	(0.142)	(0.142)	(0.142)	(0.144)	(0.144)
Role: Chief of Operations		0.001			−0.244
		(0.377)			(0.385)
Role: Fiduciary/Procurement		−0.072			−0.283
		(0.282)			(0.289)
Role: Other		0.122			0.096
		(0.231)			(0.233)
Role: Specialist		0.135			0.015
		(0.226)			(0.229)
Role: Team Leader			−0.001		−0.038
			(0.076)		(0.077)
Log Total Amount of Loan ($USD)				0.542**	0.574**
				(0.246)	(0.247)
ML Familiarity 2: Have heard of ML				0.889***	0.951***
				(0.249)	(0.253)
ML Familiarity 3: Have used ML				1.922***	2.029***
				(0.405)	(0.412)
Observations	732	732	732	732	732

Notes: Ordered logistic model coefficients shown. Dependent variable is the Likert scale response to the question: "How well did you understand the machine learning learning tool?". Larger values indicate greater perceived usefulness in the ML tool. Explainable AI treatment refers to the treatment effect of being randomly assigned into the group that received the additional screens that gave information on model performance and model explanations. *p<0.1; **p<0.05; ***p<0.01

Table 5: Explainable AI Engenders Trust Among AI Reluctant Groups

	(1)	(2)	(3)	(4)
Explainable AI ((XAI) treatment	−0.908***	−0.580***	−0.591	−2.965***
	(0.181)	(0.176)	(0.886)	(0.951)
XAI Treatment * Role: Chief of Operations	1.438***			1.412***
	(0.414)			(0.415)
XAI Treatment * Role: Fiduciary Financial Management Specialist	0.948***			1.047***
	(0.266)			(0.271)
XAI Treatment * Role: Operational Analyst	1.293***			1.262***
	(0.196)			(0.197)
XAI Treatment * Role: Other	1.083***			0.915***
	(0.219)			(0.222)
XAI Treatment * Role: Procurement Fiduciary Specialist	1.102***			0.891***
	(0.236)			(0.238)
XAI Treatment * Role: Specialist	0.246			0.045
	(0.357)			(0.359)
XAI Treatment * Role: Team Leader	0.831***			0.828***
	(0.211)			(0.212)
Log Total Amount of Loan ($USD)		1.000***		1.149***
		(0.186)		(0.193)
XAI Treatment * ML Familiarity 1-2: None/have only heard of ML		0.190		0.377*
		(0.197)		(0.203)
XAI Treatment * ML Familiarity 3: Have used ML			0.040	0.069
			(0.049)	(0.051)
Observations	732	732	732	732
Log Likelihood	−3,158.255	−3,286.368	−3,334.493	−3,069.823
Akaike Inf. Crit.	6,348.511	6,584.736	6,676.987	6,183.645

Notes: Poisson model estimates shown. Explainable AI treatment refers to the treatment effect of being randomly assigned into the group that received the additional screens that gave information on model performance and model explanations. *p<0.1; **p<0.05; ***p<0.01

Table 6: Alternative Specifications: Impact of XAI on Trust

Explainable AI treatment	0.207^{***}	0.102	0.212^{***}	0.216^{***}	0.117
	(0.063)	(0.068)	(0.063)	(0.064)	(0.372)
Role: Chief of Operations		-1.678^{***}			
		(0.211)			
Role: Fiduciary/Procurement		0.113			
		(0.117)			
Role: Operational Analyst		-0.154^{*}			
		(0.093)			
Role: Team Leader		-0.728^{***}			
		(0.094)			
Log Total Amount of Loan ($USD)		0.266			
		(0.164)			
ML Familiarity 2: Have heard of ML		-0.364^{***}			
		(0.111)			
ML Familiarity 3: Have used ML		-0.907^{***}			
		(0.117)			
ML Familiarity 4-5: Very/Extremely familiar with ML		-1.172^{***}			
		(0.168)			
Second project for user			0.301^{***}		
			(0.078)		
Difference between ML prediction and planned timespan				0.005	
				(0.008)	
Difference guess for typical delay estimate and historical delay				−0.006	
				(0.004)	
Constant	-0.963^{***}	−4.770	-1.215^{***}	-1.129^{***}	-2.160^{***}
	(0.190)	(2.934)	(0.203)	(0.427)	(0.387)
Loan Fixed Effect	Yes	Yes	Yes	Yes	No
User Fixed Effect	No	No	No	No	Yes
Observations	732	732	732	732	732
Log Likelihood	−2,115.374	−1,953.699	−2,107.932	−2,113.948	−1,647.108
Akaike Inf. Crit.	4,236.748	3,929.398	4,223.865	4,237.895	3,300.217
Bayesian Inf. Crit.	4,250.535	3,979.951	4,242.248	4,260.874	3,314.004

Notes: Poisson model estimates shown. Dependent variable is the absolute value of the change in loan delay estimate after viewing ML predictions. Larger values indicate a greater willingness to trust the ML model's predictions relative to the employees first estimate, assessed prior to using the ML tool. Explainable AI treatment refers to the treatment effect of being randomly assigned into the group that received the additional screens that gave information on model performance and model explanations. *p<0.1; **p<0.05; ***p<0.01

Online Appendix

A Additional Tables and Figures

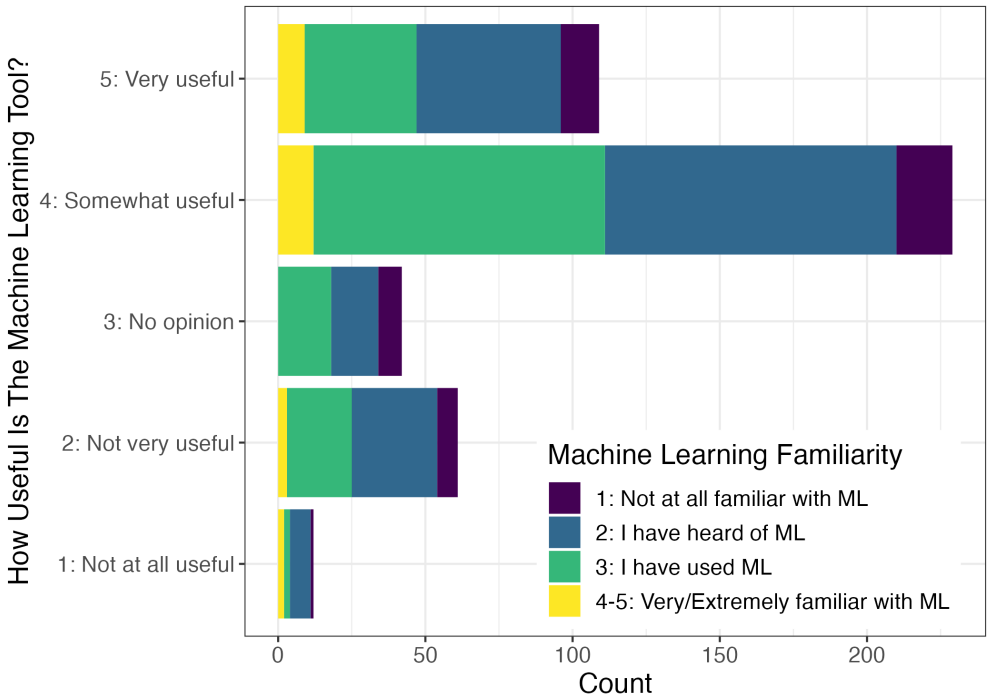


Figure A1: User-reported usefulness of the machine learning dashboard.

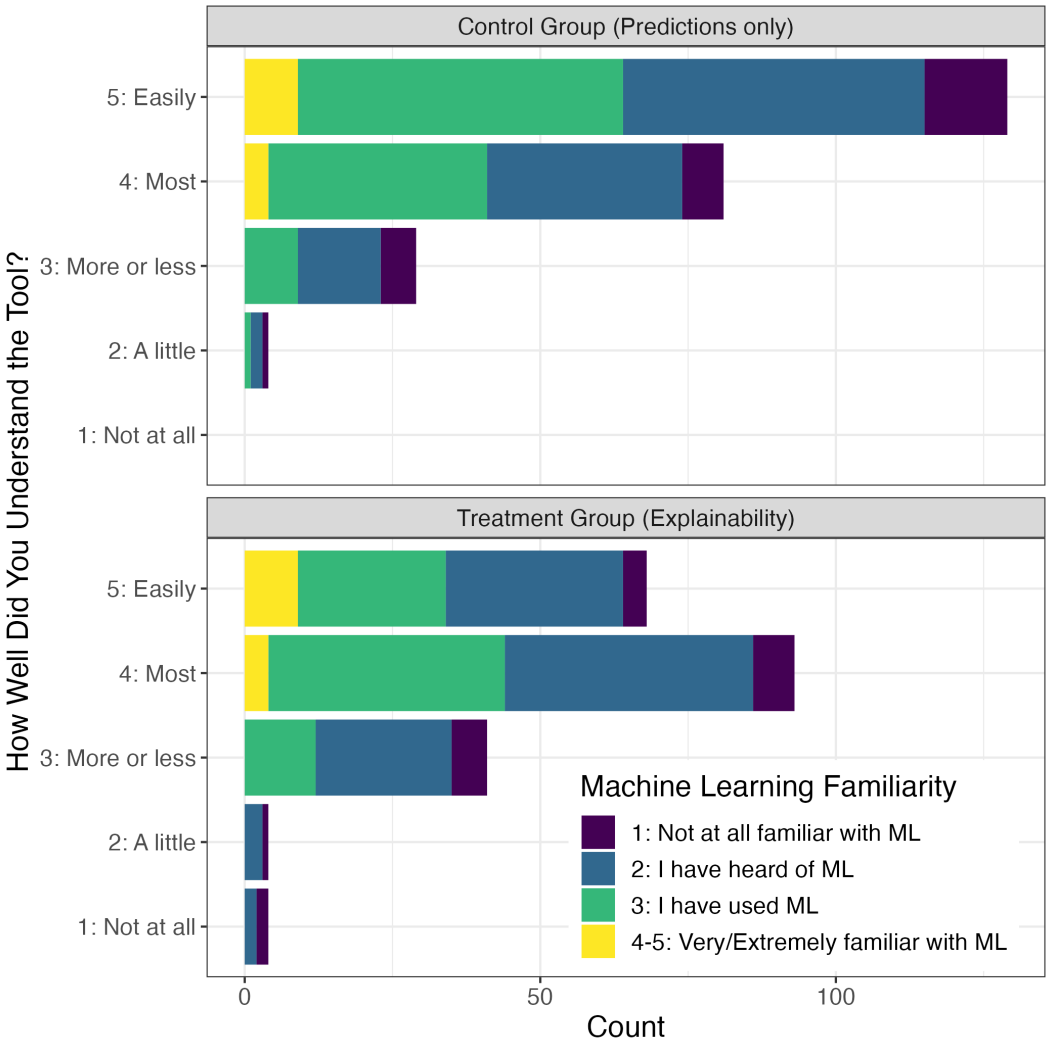


Figure A2: User-reported understanding of the machine learning dashboard, with and without the Explainable AI Component.

B Details on Machine Learning Model to Predict Loan Delays

We experimented with a variety of machine learning models to predict delay, including Random Forests, Ridge and Lasso models, and Extreme Gradient Boosted trees. The model we eventually used was a Random Forest model comprised of 1000 bootstrapped samples. We use cross validation to experiment with the optimal model hyperparameters. We cross-validated over the number of trees and the minimum node size. The analysis was performed in R programming language using the machine learning library caret.

B.1 Training Data Used to Estimate Machine Learning Model

The historical data on loan characteristics and their resulting delay status was collected between a period of 2008-2020. These comprised 518 loans (around 35,000 month-project observations). We divided the dataset into a 85% training and 15% training samples, with the intention of estimating the model on the training portion of the data and evaluating the model against the testing set. Our unit of observation is at the loan-month level, meaning for any loan i at time t there was temporal dependence for subsequent periods $t+1$. Therefore the testing and training split was performed at the loan level, with roughly 440 loans (around 30,000 month-project observations) being used to train the model, and 78 loans (around 10,000 month-project observations) being used to evaluate the model.

B.2 Model Performance

The final model had an mean average error of 11.2 months and an R^2 of 0.55 in the test set. Figure A3 presents the predicted and true months of delay a loan experienced for observations in the testing set. Each dot presented is shaded according to the distribution of underlying data, since the observations tend to be clustered around 25 months for the predicted and true values. We see that most observations lie near or close to the 45 degree line.

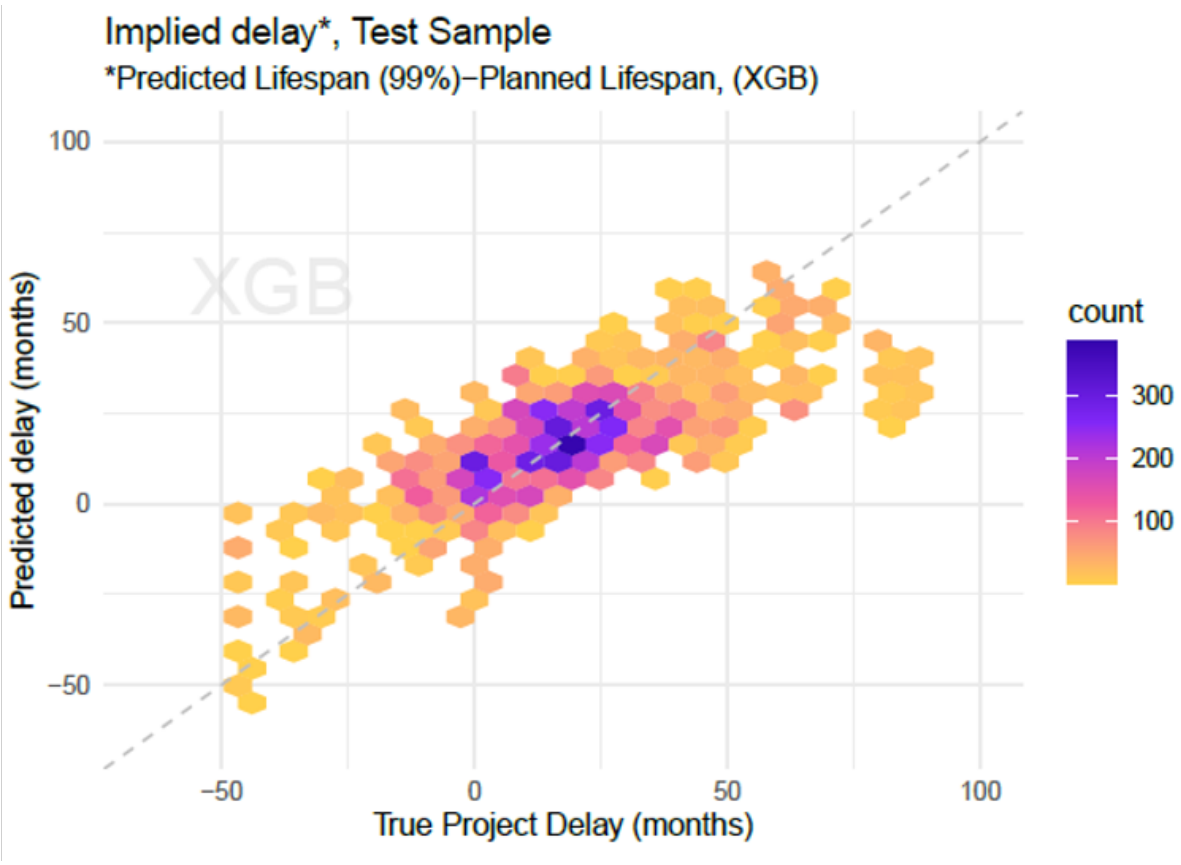


Figure A3: True loan delay and model predicted loan delay are presented for observations in the test set.

B.3 Machine Learning Model Features

The input data matrix (X_{it}) includes over 100 features consisting of project-level information (i) at age t in months since approval. Features include both fixed and temporal variables. In feature engineering to predict loan delay we wanted to capture leading indicators that would indicate a project was delayed.

The following are examples of project-level fixed variables: country, department, year and quarter of project approval, months spent in preparation, approval amount (US current dollars), project modality (type of investment loan), environmental and social safeguards risk classification, type of executing agency (central government agency, state, municipality, or other), and whether or not the project is accompanied by technical cooperation from the bank.

Examples of project-level temporal variables include: the percent of total project funds disbursed by the bank to the executing agency, the time remaining until the orig-

inal project deadline, the number of changes in the project team leader, the experience of the project team leader (cumulative years and projects managed), the number of mentions of "delays" in team leader comments on the project in a biannual review, the bank's elapsed time working with the client executing agency, and a standardized score of project results.

We also include the following annual country-level economic data, imputing missing values: GDP per capita annual growth rate, unemployment rate, and inflation.

A full list of the features included in the model are presented below.

- Basic project information
  - Country
  - Department
  - Approval year
  - Approval quarter
  - Approved amount ($USD)
  - ESG classification
  - Modality
  - Approval procedure
  - Country requires ratification
- Basic project information
  - Time spans (months)
  - Time before project start
  - Start time to ERM
  - Time approval to eligibility
  - Time eligibility to first disbursement
  - Number of missing date fields
  - Total cost of preparation
  - Time and labor hours in preparation
- Country data
  - Country has OPC TC
  - Country part of credit line
  - Country part of sequence
  - Number of loans in sequence
  - Number loan contracts
  - Country GDP growth rate
  - Country unemployment rate
  - Country inflation rate
- Executive agency experience
  - Type of agency
  - Driving distance project to country office
  - Number of projects currently managing
  - Number of projects has managed
  - Years been a client
- Project team
  - Number changes in team lead
  - Number of project team leader managing
  - Number of project leader managed in past

  - Years experience team leader
  -
- Yearly updates file
  - Language
  - Number of characters used
  - Number of fields entered
  - Key words and bigrams associated with delays
- Current loan information
- Fraction already disbursed
- Fraction disbursed relative to country IDB averages

C Survey on Machine Learning and XAI Effectiveness

All participants in the study received an email invitation to participate in the survey. The survey was available in English and Spanish and included the following questions:

Questionnaire (page 1)

The purpose of this survey is to test the utility of a new tool that predicts project delays using "machine learning" and historic data of closed IDB projects.

Please complete this survey individually. Individual results will not be shared with managers or other team members. We will use the aggregate results to understand if the new tool is useful for different audiences and how it can be improved.

---

1. What is your username?
2. What is your age range?
   - ○ Under 18
   - ○ 18-24
   - ○ 25-34
   - ○ 35-44
   - ○ 45-54
   - ○ 55-64
   - ○ 65+
   - ○ Prefer not to answer
3. Are you assigned to work in a country office or headquarters?
   - ○ Headquarters
   - ○ Country office
   - ○ Other (please specify)
4. How familiar are you with "machine learning" tools and techniques?
   - ○ **1 (Not at all familiar)**: I have never used any "machine learning" tools and/or I know very little about them.
   - ○ **2 (Not so familiar)**: I have heard of "machine learning" tools and I understand the general idea, but I've never learned any of the specific techniques involved.
   - ○ **3 (Somewhat familiar)** - I have used "machine learning" tools and/or I have taken short courses on this topic.
   - ○ **4 (Very familiar)**: I have built my own simple "machine learning" tools and/or I have taken intermediate courses on this topic.
   - ○ **5 (Extremely familiar)** : I have given trainings/workshops or tutored others on "machine learning" techniques.

Questionnaire (page 2)

Project delays

Throughout this survey, we will ask you about your perception of delays of SG investment loan projects.

1. What is your estimate (guess) of the average number of months a typical SG investment loan project is delayed?
("Delay" Date of final disbursement to executing agency - Original Disbursement Expiration Date)

*Note: On the slider below, **negative** (-) numbers mean the project finishes **early**, while **positive** numbers mean the project is **delayed**.*

-50 months (EARLY) | 0 months (ON TIME) | 50 months (DELAYED)

2. Please describe any factors influencing your guess (optional).

Questionnaire (page 3)

Projects you are working on

In this section, we will ask you about SG investment loan projects in execution that you help manage or you work on.

1. How many SG investment loan projects in execution are you currently working on? *(either directly as a team member or indirectly as an analyst)*

- ○ 0
- ○ 1
- ○ 2 or more

{If respondent selects "0", survey ends.}

Pages 4-6 loop one or two times, depending on the number of projects selected in page 3.

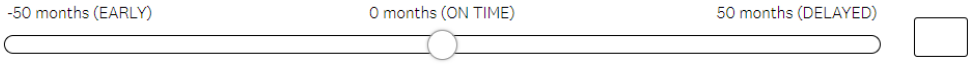
Questionnaire (page 4)

You selected {1 or 2} project(s).

We will ask you questions about these projects. If you work on more than 2 projects, please select the 2 projects you are most familiar with.

1. What is the {first/second} project number?
2. What is your role on this project?
   - ○ Team Leader
   - ○ Alternate Team Leader
   - ○ Operational Analyst
   - ○ Chief of Operations
   - ○ Fiduciary Financial Management Specialist
   - ○ Procurement Fiduciary Specialist
   - ○ Specialist
   - ○ Other team member (please specify):
3. If you had to guess, how many months do you think this project will be delayed relative to the original disbursement expiration date?
   ("Delay" = Date of final disbursement to executing agency - Original Disbursement Expiration Date)

*Note: On the slider below, **negative** (-) numbers mean the project finishes **early**, while **positive** numbers mean the project is **delayed**.*

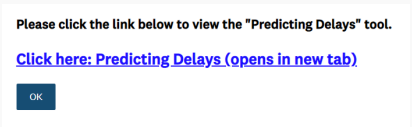


Questionnaire (page 5)

Video: Predicting Delays Tool

{Respondents are prompted to watch a video that explains how to use the tool. The video for the Control Group is 4 minutes, while the video for the Treatment Group is 10 minutes.}

{Next, respondents are prompted to view the tool using a link. Control Group respondents have a private link for the tool with a single tab, while the Treatment Group respondents have a private link for the tool with two additional tabs.}



1. Were you able to open the tool?
   - ○ Yes
   - ○ No

{If respondent selects "No", survey ends.}

Questionnaire (page 6a)

Find your project in the tool

Please try to find your project in the tool.

1. Were you able to find your project in the tool?

   ○ Yes

   ○ No

2. After viewing this tool, did you change your guess of this project's delay?

   ○ Yes

   ○ No

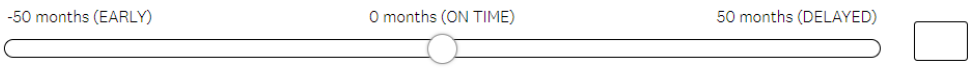
{If respondent selects "Yes", page 6b starts, otherwise skip to page 7.}

Questionnaire (page 6b)

New guess

1. After viewing the tool, what is your new guess for the number of months this project will be delayed?
("Delay"= Date of final disbursement to executing agency - Original Disbursement Expiration Date)

Note: On the slider below, ***negative*** *(-) numbers mean the project finishes* ***early****, while* ***positive*** *numbers mean the project is* ***delayed****.*



2. What factors influenced your decision to update your guess (optional)

Questionnaire (page 7)

Reaction to tool

1. How useful is the tool for your work? (Please comment)
   - ○ ⋆ Not at all useful
   - ○ ⋆⋆ Not very useful
   - ○ ⋆⋆⋆ No opinion
   - ○ ⋆⋆⋆⋆ Somewhat useful
   - ○ ⋆⋆⋆⋆⋆ Very useful
2. How well did you understand how to use the tool? (Please comment)
   - ○ ⋆ Could not understand at all
   - ○ ⋆⋆ Only understood a little bit
   - ○ ⋆⋆⋆ Understood more or less
   - ○ ⋆⋆⋆⋆ Understood for the most part
   - ○ ⋆⋆⋆⋆⋆ Understood easily
3. Do you think this tool is innovative? (Please comment)
   - ○ ⋆ Not at all
   - ○ ⋆⋆ Somewhat
   - ○ ⋆⋆⋆ No opinion
   - ○ ⋆⋆⋆⋆ Fairly innovative
   - ○ ⋆⋆⋆⋆⋆ Very innovative
4. How would you improve the tool? (optional)