Explainable AI and Adoption of Financial Algorithmic Advisors: an Experimental Study

Daniel Ben David
The Hebrew University of Jerusalem
Intuit Inc.

Talia Tron
Intuit Inc.

Yehezkel S. Resheff
Holon Institute of Technology

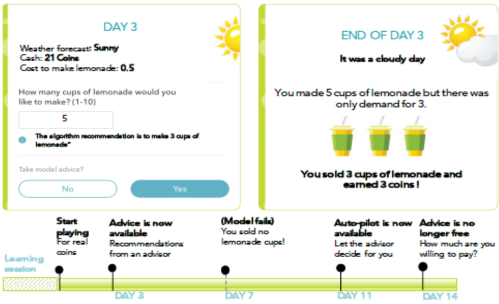
Abstract

We study whether receiving advice from either a human or algorithmic advisor, accompanied by five types of Local and Global explanation labelings, has an effect on the readiness to adopt, willingness to pay, and trust in a financial AI consultant. We compare the differences over time and in various key situations using a unique experimental framework where participants play a web-based game with real monetary consequences. We observed that accuracy-based explanations of the model in initial phases leads to higher adoption rates. When the performance of the model is immaculate, there is less importance associated with the kind of explanation for adoption. Using more elaborate feature-based or accuracy-based explanations helps substantially in reducing the adoption drop upon model failure. Furthermore, using an autopilot increases adoption significantly. Participants assigned to the AI-labeled advice with explanations were willing to pay more for the advice than the AI-labeled advice with "No-explanation" alternative. These results add to the literature on the importance of XAI for algorithmic adoption and trust.

Introduction

When algorithms take part in decision making with significant impact to individuals, there is a burden of explainability that naturally falls on the providers of the system. In some cases there are regulations that imposes obligatory explanation of decisions as an integral part of their output. However, in many other cases, and in various fields, the importance of explanations is first and foremost in the effect on the perceived trustworthiness of the system, and hence on the readiness of consumers to adopt (RTA) the AI service, and pay for it. The first important attribute when considering what explanation to generate is the type of information the explanation should convey. We find in the literature three main approaches to explanations: (1) Global Explanations, (2) Local explanations and (3) what might be called Social Influence Explanations. The aim of this type of explanationis to convey to a human what the algorithm is doing rather than explain the process that lead to a specific prediction or decision. In our study, we attempt to form a synergy between explainable AI and the multiple fields that have previously studied algorithmic adoption, and machine-human relations. We use the quantitative measure of Readiness to Adopt (RTA), simply defined as the fraction of users that use the AI system when presented with the choice, and later the Willingness to Pay (WTP) to explore acceptance and its relations to trust and user satisfaction. We study the impact of the explanation type (both global and local) on the adoption of and payment willingness towards an AI financial decision-making advisor. We do so in a unique experimental framework, in a controlled environment with real money consequences.

Experimental Design


DAY 3

Weather forecast: Sunny
Cash: 21 Coins
Cost to make lemonade: 0.5

How many cups of lemonade would you like to make? (1-10)

5

The algorithm recommendation is to make 3 cups of lemonade"

Take model advice?

No Yes

END OF DAY 3

It was a cloudy day

You made 5 cups of lemonade but there was only demand for 3.

You sold 3 cups of lemonade and earned 3 coins !

Learning session
Start playing
For real coins
Advice is now available
Recommendations from an advisor
(Model fails)
You sold no lemonade cups!
Auto-pilot is now available
Let the advisor decide for you
Advice is no longer free
How much are you willing to pay?
DAY 3
DAY 7
DAY 11
DAY 14

The study consistent of a 3 parts as follow:

Pre-game quantitative questionnaire. Participants had a time limit of 3 minutes to answer 3 simple mathematical questions (addition and subtraction). Upon completing this part, participants earned 20 initial game coins that were used in the main part of the study. Each game coin during the game was worth 2 U.S. cents. The apurpose of this stage was twofold: The first was to ensure participants' attention during the experiment. The second relates to the mental accounting literature. We wanted participants to treat the game coins account as money that they earned while investing effort, rather than money that they obtained as a "reward".

The main part of the study. A fun, interactive, decision-making game - The Lemonade Stand. In this part participants could gain more coins depending on their decisions, and potentially use an advisor, that was labeled differently with several explanation (five different conditions), in doing so. We evaluate adoption over time and in different situations (**see game-flow description and decision illustration figure**).

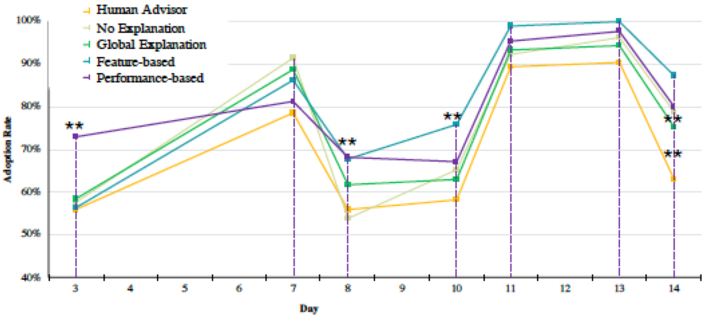
A post-game questionnaires about trust, engagement, explanation satisfaction and personal demographic details.

Experimental Conditions:

Condition	Explanation Type
Human Expert	The human advisor recommendation is to make 6 cups of lemonade.
No explanation	The algorithm recommendation is to make 6 cups of lemonade.
Global explanation	Based on data from lemonade stands over several years, the algorithm recommendation is to make 6 cups of lemonade.
Feature-based	Based on data from lemonade stands over several years, your previous sales, and market demand, the algorithm recommendation is to make 6 cups of lemonade.
Performance-Based	Based on data from lemonade stands over several years, the algorithm recommendation, with 90 percent certainty, is to make 6 cups

Main Results

We compare the Readiness to Adopt (RTA) for the different experimental condition groups throughout the experiment "two weeks" game duration. First, we examine the effect of human versus algorithmic origin of advice with no further explanation. Results show no evidence of algorithm aversion. Next, we considered the effect of the type of explanations provided by the algorithmic advisor on RTA during the different phases of the game. Attaching the accuracy of the model to the explanation created a significant difference and created the highest initial adoption rates, with RTA of 73% for the Accuracy compared to an RTA average of 57% for the other explanation types. During the first four days, the advice provided in all experimental conditions allowed participants to make the highest possible profit (i.e. perfect predictions and advice). Results suggest that when advice performance is immaculate, the type of explanation presented with it is less important to individuals. When advice fails, during day 8, we document that the Feature-Based and Performance-Based have significantly lower adoption drop compare to the alternative with No explanation. In addition, when the trust was perceived as higher, it lowered the drop rates and increased the recovery from the failure. On day 11, when subjects had the option to select auto advice, we observe adoption jump of 40% on average to a mean RTA of 95% on day 11 and up to 97% on day 13. On the last day we introduced a new situation where the advisor is no longer free and asked participants whether they want the advice and if so, how match they are willing to pay for it. During this round we observe adoption drop of 21% for the algorithmic alternatives. This was despite the fact that we allowed the subjects to self-determine the value (the price) of the advice. We find that participants assigned to the No-explanation alternative were willing to pay the least for the use of the advice with mean payment of 1.005 game coins, compared to a mean payment of 1.774 game coins for all other "AI advice" alternatives that includes explanations A 76.5% gap.



Contribution

(1) First, we document that there is no single best explanation that fits all. The type of explanation that is best suited to promote trust is time and situation dependent. Namely, the "What" we should explain depends on the "when". Interestingly, participants were more inclined overall to adopt AI-given advice than follow a human expert. The effect of the model (and human advice) failure on subsequent advice taking is remarkable, with a single failure causing adoption rates to plunge.

(2) Second, our results suggest that often the end-user doesn't need to know more than very general facts to accept the system. Our study shows that such general explanations yield good results in terms of RTA and WTP compared to the alternative of no explanations, and especially after AI failure. Moreover, stating accuracy statistics about the algorithm can further strengthen the aforementioned effect. Furthermore, explanations increased the participants willingness to pay for the advice. Results of the current study highlight the potential utility of simple explainability solutions in these aspects.

(3) Third, our experimental paradigm of the lemonade stand game applies the RTA measurement in different and evolving situations into the explainable AI literature. This framework may be utilized to answer other questions in the field of explainable AI and trust in AI.

CONCLUSION

To the best of our knowledge, this is the first study to directly evaluate the behavior of users in response to varying types of textual global and local explanations of AI over time in different situations. Our experimental paradigm allows testing initial adoption in several experimental conditions (defined by the different types of explanations), as well as the evolution of the relations over time when the AI advice proves to be useful, or after it fails. Post-game questionnaires further allow integration of measures common in the behavioral sciences for additional validation of our findings.