informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *Management Science*

ISSN 0025-1909, EISSN 1526-5501

Algorithmic Transparency and Portfolio Choices

Béatrice Boulu-Reshef
CY Cergy Paris Université, THEMA, CNRS, beatrice.boulu-reshef@cyu.fr

Alexis Direr
Université d'Orléans, LÉO, University of Orléans, alexis.direr@univ-orleans.fr

Mehdi Louafi
Université d'Orléans, LÉO, mehdi.louafi@univ-orleans.fr

Abstract. We test whether profile-based explanations affect adherence to algorithmic risk recommendations on a leading French robo-advisor. In a randomized controlled trial embedded in the platform's interface ($N = 4,645$ saving-plan contracts), users were randomized to receive graphical explanations of the factors underpinning their recommended risk score or no additional information. Explanations do not raise acceptance of the recommended profile. Device context matters: desktop users follow recommendations less than phone users, and among desktop users who deviate downward, explanations lead them to deviate further from the recommendation. In a real-market setting, profile-based transparency does not deliver the expected compliance gains; while informative, it is not a universal lever for adherence and should be adapted to the usage context.

Funding: This research has been conducted within the PREF research initiative under the aegis of the Europlace Institut of Finance, a joint initiative by YOMONI and was supported by the APR-IA "AcceptAlgo" funded by the Centre-Val de Loire Region and the ANR "AESOP".

Key words: Robo-Advisor, Financial advice, Portfolio Choices, Household Finance, Algorithmic Transparency

- Expressions trop journalistiques
- Ressortir engagement avec cluster
- Raconter dans engagement pc + modif events et open events
- Conclusion : engagement pc raccorder avec histoire

1. Introduction

Robo-advisors —algorithmic, largely automated portfolio advice delivered via digital platforms— promise scalable, low-cost guidance (Abraham et al. 2019, D'acunto and Rossi 2022), consistent with broader FinTech-driven declines in intermediation costs (Philippon 2019), yet adoption remains modest (**?**). A frequently cited friction is algorithmic transparency: recommendations are perceived as "black-box," users lack visibility into why a given risk profile fits their circumstances, and the loss of human interaction can erode trust (Morana et al. 2020, Jung et al. 2018). Explainable AI (XAI) proposes user-facing explanations to address this gap; yet, there is limited field evidence on whether explanations actually change behavior in high-stakes financial decisions (Adadi and Berrada 2018, Anjomshoae et al. 2019, Arrieta et al. 2020, Riedl 2019, Weitz et al. 2021). We therefore ask whether profile-based explanations, concise disclosures of the drivers of a recommended risk score benchmarked to similar users, increase acceptance of algorithmic risk profile recommendations on a market-deployed robo-advisor.

We study this question with a randomized controlled trial embedded in the interface of a leading French robo-advisor. The experiment randomized 4,645 saving-plan contracts to one of two conditions: (i) a treatment in which users received graphical, profile-based explanations detailing the relative importance of variables (economic situation, liquidity needs, investment horizon, demographics, risk tolerance) and benchmarking them to a "typical similar user"; or (ii) a control without any additional information beyond the recommended risk profile. Our primary outcome is acceptance, defined as choosing the recommended risk profile at the decision point. We also track the exact deviation from the recommended profile and engagement metrics. Sessions are observed on desktop and mobile, allowing us to study how the usage context moderates responses.

Three results stand out. First, explanations do not raise acceptance: treatment and control accept at statistically indistinguishable rates. Second, device context matters: desktop sessions exhibit lower baseline adherence than mobile. Third, among desktop sessions that choose a safer-than-recommended portfolio, explanations increase the magnitude of the downward deviation. Beyond this subgroup, we find little systematic heterogeneity and no improvement in engagement. Taken together, profile-based transparency is not a universal compliance tool and could produce counter-intuitive outcomes.

Our paper makes three contributions. First, we provide in-market experimental evidence on a widely advocated XAI intervention, and we document no effect on the primary outcome and a counterintuitive effect for a meaningful subgroup. This helps bridge a literature heavy on technical explainability with behavioral responses in high-stakes, real transactions. Second, we show that device context—desktop vs. mobile sessions—systematically moderates adherence and shapes how explanations operate. Third, we offer managerial guidance: for deployments seeking higher compliance, adding profile-based explanations does not guarantee higher adherence. Firms should test context-contingent designs (e.g., lighter explanations on desktop at the acceptance step, progressive disclosure, alternative benchmarks) rather than expect uniform gains from transparency.

We implement a single-feature intervention by mirroring how a human advisor would justify a recommendation, grouping drivers into meaningful categories and situating the user in a peer reference class, while preserving the product's decision flow. The randomization was implemented within the production system and analyzed independently by the authors.

Our findings speak to two ongoing debates. First, they nuance the view that more algorithmic transparency automatically builds trust and uptake: transparency can influence deviations from the recommendation instead of the decision to accept it. Second, they highlight that interface constraints are not second-order. Explanations that are beneficial in mobile, attention-limited contexts may be counterproductive on desktop, where users may more easily reveal a preference and potentially misinterpret the explanation.

The remainder of the paper proceeds as follows. Section 2 situates our study within research on explainability, robo-advising, and digital finance. Section 3 details the experimental design and measurement. Section 4 reports the main effects and robustness. Section 5 concludes with implications for the design of explainable robo-advice.

2. Related work

2.1. Algorithmic transparency and explanations

Concerns about the opacity of modern machine learning (ML) and Artificial Intelligence (AI) systems have fueled calls for algorithmic transparency from stakeholders, regulators,[1] and users (Castelvecchi 2016, Preece et al. 2018). XAI responds by making system outputs comprehensible to humans (Adadi and Berrada 2018), with two complementary aims: (i) clarifying the drivers of individual or aggregate predictions (Anjomshoae et al. 2019, Arrieta et al. 2020, Beaudouin et al. 2020) and (ii) supporting understanding, acceptance, and trust (Cheng et al. 2019, Cai et al. 2019, Shin 2021, Weitz et al. 2021).

On the methods side, XAI distinguishes inherently interpretable models from black boxes equipped with post-hoc explanations (Arrieta et al. 2020). Explanations can be local, targeting a single decision (e.g., LIME (Ribeiro et al. 2016)), or global, describing overall feature influence (e.g., SHAP/SAGE (Lundberg and Lee 2017, Cohen et al. 2005, Covert et al. 2021); see also Speith 2022, Krzyziński et al. 2023). A common design is feature-importance attribution and short rationales ("recommended because of A and B; despite C"), widely used in recommender systems (Nunes and Jannach 2017).

On the human side, experiments in Human-Computer Interaction and Computer Science show that explanations typically improve understanding, but their effects on confidence and acceptance vary with context and presentation: example-based or agent-mediated explanations can help, yet gains in comprehension do not always translate into greater trust or adoption (Cai et al. 2019, Shin 2021, Weitz et al. 2021, Rader et al. 2018, Eslami et al. 2018, Herlocker et al. 2000, Cramer et al. 2008). Human-centered work stresses that "good" explanations must anticipate edge cases and align with user goals (Riedl 2019). In sum, the literature offers mature tools for transparency but mixed behavioral predictions, motivating field evidence in high-stakes settings.

2.2. XAI, explanations, and robo-advisors

In robo-advising, opacity has been found to have an impact on trust and delegation (Patel and Lincoln 2019, Board 2017, Bianchi and Brière 2020). Users often display algorithm aversion, thus preferring human judgment even when algorithms perform better (Dietvorst et al. 2015, 2018). Lab and online studies suggest that transparent explanations can mitigate aversion and support adoption, though effects depend on format and timing: accuracy or feature-based disclosures can raise uptake and preserve trust after errors (Ben David et al. 2021); SHAP-style visualizations that explicitly link inputs to recommendations improve comprehension and engagement when kept simple (Deo and Sontakke 2021). Related finance models highlight the risks of over-personalization and the role of transparent human–AI interaction (Capponi et al. 2022, Bianchi and Briere 2021), while policy work emphasizes accountability and explainability in automated advice (Fein 2017, Strzelczyk 2017, European Commission 2019). Overall, the literature posits transparency as a lever for trust and delegation, but it remains an empirical question whether profile-based disclosures change behavior in market settings.

2.3. Digital finance and the decision context

Digital finance reshapes how households decide, not just what they can access. By embedding decision support, reducing frictions, and shifting interactions onto smartphones, platforms alter attention, search, and default uptake. Empirically, broader digital access is associated with greater participation and risk-taking: exposure to digital services correlates with higher allocation to risky assets and reduced risk aversion (Hong et al. 2020); digital finance increases the extensive margin of risky-asset holding (Shen et al. 2022). In China, a one-percent increase in digital finance (DFII) corresponds to a 0.13-percent rise in the share of risky assets and raises the likelihood of holding any risky assets (Hu et al. 2024). The interaction channel is therefore not neutral for portfolio choices.

Behaviorally, the mobile mindset favors fast, heuristic processing over deliberation (Lurie et al. 2018, Kahneman 2013). In a lab setting with device randomization, smartphone users display a present-bias parameter about six percentage points lower and are roughly twelve points less willing to pay search costs for payoff-relevant information (Mograbi 2022); habitual phone use similarly increases reliance on defaults and recommendations relative to larger screens (Wang et al. 2015). For explanation design, this implies that identical transparency features—such as profile-based disclosures of drivers and peer benchmarks can raise comprehension yet have device-contingent effects on acceptance: mobile may reinforce default adherence, whereas desktop affords scrutiny and adjustment. This motivates our focus on desktop versus mobile sessions and frames our interpretation of heterogeneous impacts in the experiment.

3. Experimental design

3.1. Onboarding flow ("tunnel")

In our setting, each prospective client who engages with the robo-advisor completes a standardized, multi-step onboarding process. For consistency with the partner's internal terminology[2], we refer to this flow as the "tunnel" when discussing specific screens and steps. This structured sequence consists of four distinct stages, referred to as "Steps", through which clients progress sequentially. Each step serves a specific function in the investment decision-making and account setup process (see Figure 1). In Step 1, a unique User ID is assigned, and the step concludes when the client initiates an investment simulation. In Step 2, the client completes a comprehensive financial questionnaire, which serves as the foundation for the risk assessment and portfolio recommendation process. Submission of this questionnaire simultaneously generates a new Contract ID linked to the corresponding User ID. Step 3 comprises the validation of the proposed risk profile, investment vehicle selection, and portfolio allocation against the client's stated objectives and constraints. At Step 3, users may open a "Modify" panel listing alternative risk profiles with side-by-side allocations. Acceptance is recorded immediately after closing this panel when the final profile is confirmed and matches the originally recommended profile. Finally, Step 4 requires submission of all requisite legal documentation and the formal execution of the contractual agreement, thereby completing account opening with the robo-advisor.

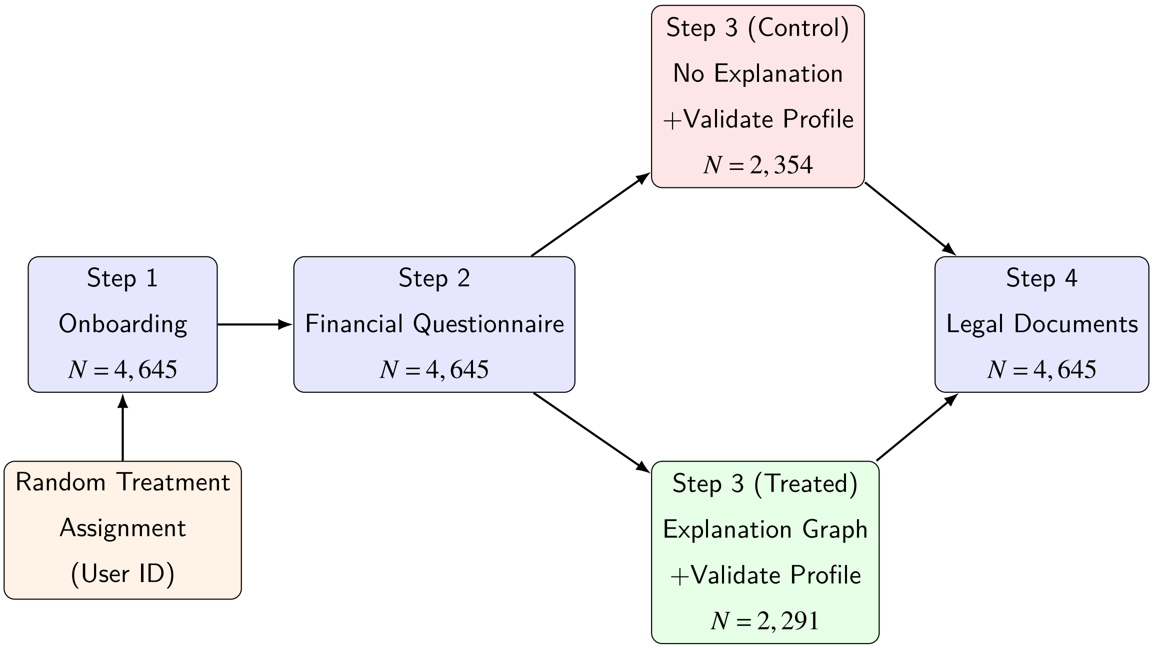


Figure 1 Flowchart of the Experiment

Table 1 Target asset allocation by risk profile (in %)

	Risk Profiles								
Asset Class	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Euro Funds	70	60	40	20	0	0	0	0	0
Bond ETFs	15	20	30	40	50	40	30	20	0
Stock ETFs	15	20	30	40	50	60	70	80	100

A key component of this onboarding process is the financial questionnaire, which gathers comprehensive information regarding the client's financial situation, investment preferences, and constraints. Specifically, the questionnaire includes inquiries on liquidity needs, risk tolerance, financial literacy, investment horizon, and investment objectives. Additionally, it collects various economic indicators such as financial wealth, income levels, savings capacity, and the amount allocated for the initial investment. Furthermore, socio-demographic variables are also recorded to provide a more holistic understanding of the investor profile. Upon completion of the questionnaire, the collected data are processed through a proprietary algorithm designed to generate a risk score. This score is instrumental in determining the client's risk profile, which subsequently dictates the composition of their recommended investment portfolio. The algorithm not only assesses the appropriate level of risk exposure but also identifies the optimal investment vehicle. For the purposes of our study, we restrict attention to a tax-efficient saving plan (*assurance-vie*), the platform's flagship product through which most clients invest. Focusing on a single wrapper ensures a common decision environment and maximizes sample size.

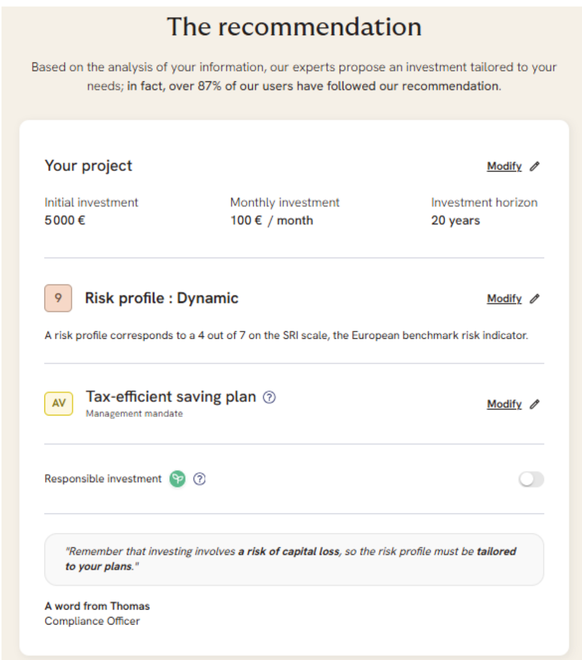
Within the context of "*assurance-vie*" products, the generated risk score ranges from 2 to 10. A risk score of 2 corresponds to a highly conservative portfolio, allocated mostly to "Fonds Euro" (see Table 1). This investment vehicle provides capital protection and is characterized by low risk, albeit with limited return potential. At the opposite end of the risk spectrum, a score of 10 represents a fully equity-based portfolio, composed entirely of stocks, thus entailing a significantly higher level of risk. Intermediate risk profiles, with scores between 2 and 10, are defined by varying allocations between "Fonds Euro", bonds, and equities. As the risk score increases, the proportion of "Fonds Euro" decreases, while the allocations to bonds and equities rise accordingly, reflecting a progressive shift towards a more aggressive investment strategy. Importantly, after receiving their assigned risk profile, prospective clients have the option to access a "modify option", which allows them to compare their recommended profile against other available risk profiles. If desired, they may adjust their final risk selection before proceeding with investment validation and contract signing.

The intervention is an A/B test with randomization at Step 1 and treatment display at Step 3, immediately after completion of the Step 2 questionnaire. The treatment group receives an explanation of the assigned risk profile, which focuses solely on the rationale behind the score generated by the algorithm. In contrast, the baseline group does not receive any explanatory information. The treatment is randomly assigned at the beginning of Step 1 based on the User ID. Consequently, if a prospective client undergoes the onboarding process multiple times, a frequent occurrence in practice, they will consistently receive the same treatment condition. This design ensures that exposure to the intervention remains stable, thereby minimizing selection bias and preserving the integrity of the experimental framework. Randomization achieved balance across pre-treatment covariates; Mann–Whitney and χ^2 tests fail to reject equality across arms (all $p > 0.11$; Bonferroni-adjusted $p \geq 0.47$, see Table 5 in Appendix B).

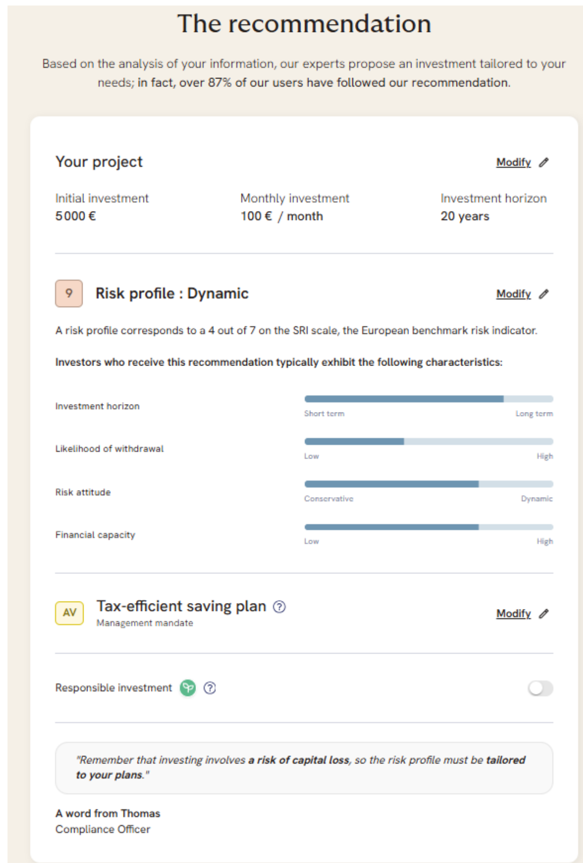
3.2. Explanation Methodology

In this study, we focus on the interpretability of recommendations generated by a financial algorithm embedded within a robo-advisor. The algorithm computes a risk profile score for prospective clients, which in turn informs the investment recommendations they receive. To improve transparency and user understanding of the profile assessment process, we implement a graphical representation that visually communicates the relative importance of various factors contributing to the profile score.

Specifically, the bar chart displays normalized weights for key determinants (liquidity constraints, risk preferences, economic characteristics, investment horizon) on a 0–10 scale. For visualization, we truncate displayed bars to the range 1–9 to avoid saturation at extremes while not altering relative rankings. An example of the visual explanation shown to treated users is provided in Figure 2a and 2b. The figure illustrates the characteristics of a user assigned risk profile 9. The accompanying bars show how individuals with similar profiles typically score on attributes such as investment horizon, withdrawal likelihood, risk attitude, and financial capacity.



(a) Recommendation Interface for Non-Treated Users, translated from French



(b) Recommendation Interface for Treated Users, translated from French

The graphical content is tailored to each client's assigned risk profile, which ranges from 2 to 10. Crucially, the explanation is profile-based: for each risk profile, the bar chart illustrates the mean normalized importance of each determinant across all users assigned that same score. In other words, it shows what characterizes the "typical" individual in that risk category, offering users comparative insights into what the algorithm considers the average user within this risk class. Displayed "importance" reflects model weights and is not a causal attribution. This approach embeds the explanation within a broader reference frame and seeks to enhance interpretability by contextualizing the recommendation with feature-importance cues linked to the recommended risk profile.

The financial algorithm underlying the robo-advisor aggregates responses from a financial questionnaire to compute a global risk score. Each response contributes to the final score based on a predefined weighting system that reflects the relative importance of various attributes in assessing a client's investment profile. Although the scoring process is non-linear—such that small changes in inputs do not always produce proportional changes in the output—it is internally transparent: the structure and logic of the model are fully observable to the platform designers. In other words, the algorithm qualifies as a "white-box" model from a backend perspective. The display is a faithful transformation of the production scoring logic with no surrogate or post-hoc explainer used. This internal transparency ensures that the graphical representation used in the intervention remains an accurate reflection of the underlying risk assessment process, allowing for a reliable interpretation of how various factors influence the final investment profile. However, this interpretability is not directly available to end users, who only see the final score without insight into how their responses shaped it. The graphical explanation provided in our intervention serves to bridge this gap by offering users an interpretable summary of the main factors influencing their assigned risk profile, thereby reflecting the logic of the underlying model in an accessible and meaningful way to prospective clients.

Our approach builds on established techniques in the XAI literature, particularly feature importance attribution methods that highlight which inputs drive model outputs. While our algorithm is not a black-box model requiring post-hoc methods like SHAP, the graphical explanation serves a similar function: to make the recommendation logic more transparent and intuitive. By delivering personalized, profile-based explanations grounded in real user profiles, our design aims to foster greater comprehension and trust in algorithmic recommendations, contributing to the broader literature on interpretable decision support in finance.

3.3. Outcome variables

We exploit a rich array of outcome measures, which we organize into two broad categories: portfolio choices, capturing how users made decisions once they receive a risk-score recommendation, and engagement outcomes, reflecting the intensity and nature of interaction with the "Tunnel" interface.

Portfolio choices focus on (i) the acceptance of the recommended risk profile; (ii) the deviation, which retains the sign of the deviation to distinguish upward from downward adjustments; and (iii) separate positive and negative deviations to gauge the intensity of upward versus downward modifications.

Engagement outcomes are inspired by the XAI and recommender-systems literatures and quantify user interactions with the platform: (i) the total number of attempts across all steps (whether or not a contract was ultimately signed); (ii) the number of opening events, defined as instances in which users clicked to access the modification interface; and (iii) the number of modification events, defined as instances in which users actively adjusted their risk profile within that interface.

Further details on the distribution of these outcome variables can be found in Table 6 in Appendix B.

3.3.1. Portfolio Choices

Acceptance of score recommendation We define a binary variable $Accept_i$ that equals 1 if user i selects exactly the risk score recommended by the robo-advisor, and 0 otherwise. This outcome measures users' willingness to follow the algorithmic suggestion.

Deviation measures from the recommended score We capture three related measures of how users adjust their chosen score relative to the robo-advisor's recommendation.

First, the Deviation D_i preserves the sign of the difference—positive for upward adjustments and negative for downward adjustments—thereby indicating the deviation direction in risk-profile notches. Formally:

$$\mathrm{D}_i = \mathrm{ChosenRP}_i - \mathrm{RecommendedRP}_i.$$

In this measure, when clients chose a score that is higher (lower) than the recommended score, the deviation is recorded using a positive (negative) value. This allows for documenting the direction of deviation from the recommended score, accounting for the sign of the deviation, if there is one.

Second and third, we subset signed deviations into separate measures of upward and downward modifications, respectively denoted PD_i and ND_i. Then, on the subsample with $D_i > 0$ we define

$$\mathrm{PD}_i = D_i \quad \big(\text{for } D_i > 0\big),$$

and, on the subsample with $D_i < 0$ we define

$$\mathrm{ND}_i = D_i \quad \big(\text{for } D_i < 0\big).$$

3.3.2. Engagement Outcomes

Total number of attempts We define $TotAttempts_i$ as the total count of times user i completed the Tunnel until Step 3, irrespective of whether a contract was ultimately signed. This aggregate metric captures overall engagement with the onboarding process.

Number of opening events We define $OpenEvents_i$ as the number of instances in which user i clicked to access the modification interface. Each opening event reflects a user's decision to review or reconsider the recommended risk profile.

Number of modification events We define $ModEvents_i$ as the number of instances in which user i actively adjusted their risk profile within the modification interface. This outcome measures the intensity of interaction with the recommendation system.

4. Results

4.1. Descriptive statistics

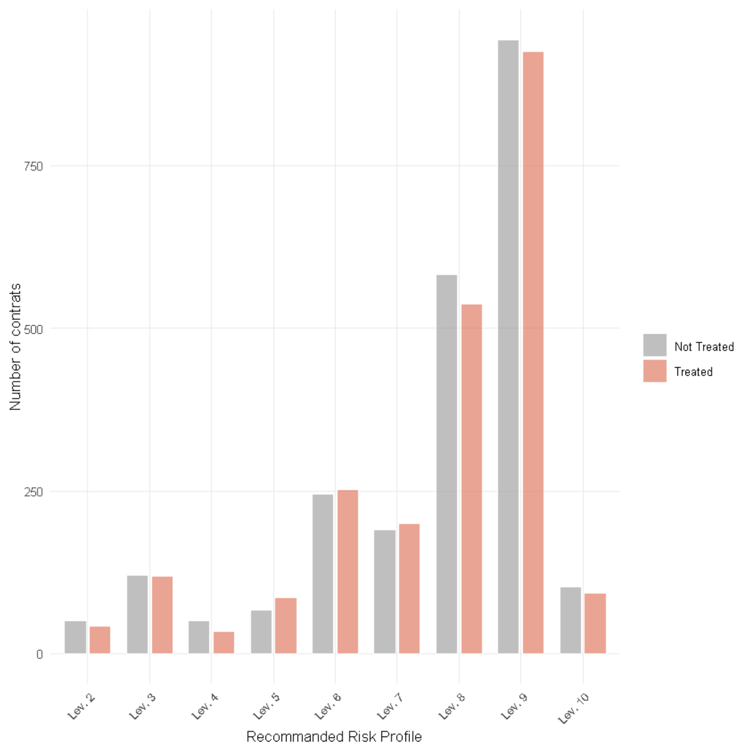


Figure 3 Distribution of Recommended Risk Profile by Experimental Condition

During the 105-day field experiment, 4,645 contracts[3] were signed, corresponding to $N = 3,856$ unique users, with an average of 1.20 contracts per user (SD = 0.57). Of these, 2,354 belonged to the control group and 2,291 to the treatment group. The mean age of users was 38.00 years (SD =

13.00 years). Annual income was distributed as follows: 25.67% reported €30,000–45,000; 20.82% €60,000–100,000; 19.73% €45,000–60,000; 19.24 % < €30,000; 10.66% €100,000–150,000; and 3.86% > €150,000. Based on financial-knowledge items, 16.8% of users were classified as "*Non initié*" (uninitiated), 43.8% as "*Néophyte*" (beginner), and 39.5% as "*Sachant*" (knowledgeable). Finally, 86.87% of users were deemed novice in financial experience, while 13.12% were deemed experienced.

Baseline balance tests reveal no significant differences between the control and treatment groups in the distribution of recommended risk profiles (two-sided Mann–Whitney U test, p-value = 0.746; see Figure 3) nor in the distribution of device sessions (personal computer vs. phone; two-sided Mann–Whitney U test, p-value = 0.614; see Figure 10 in Appendix B). Additionally, there were no differences in the distribution of user type — "first-time users", who had not used the platform before the experiment, versus "returning users", who had used the platform beforehand (two-sided Mann–Whitney U test, p-value = 0.742; see Figure 11 in Appendix B).

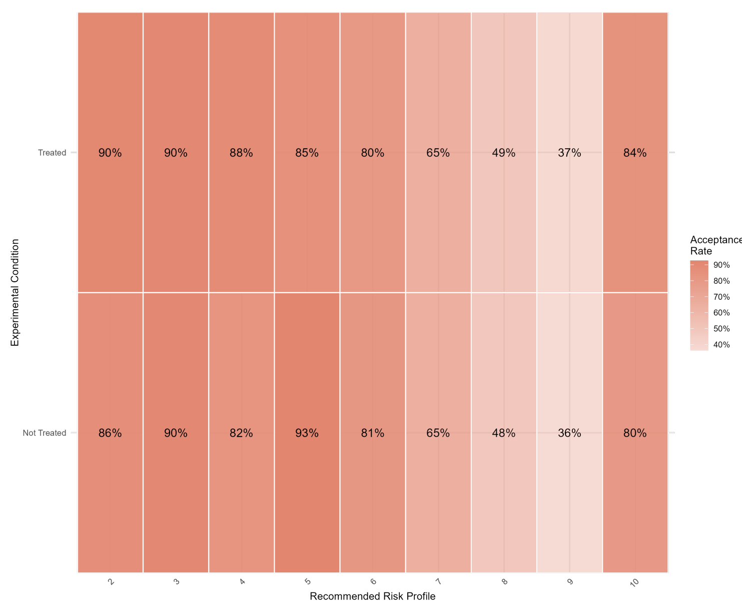
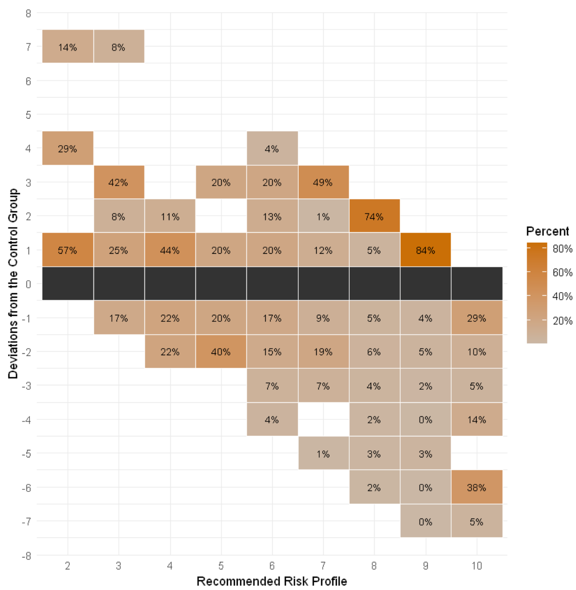


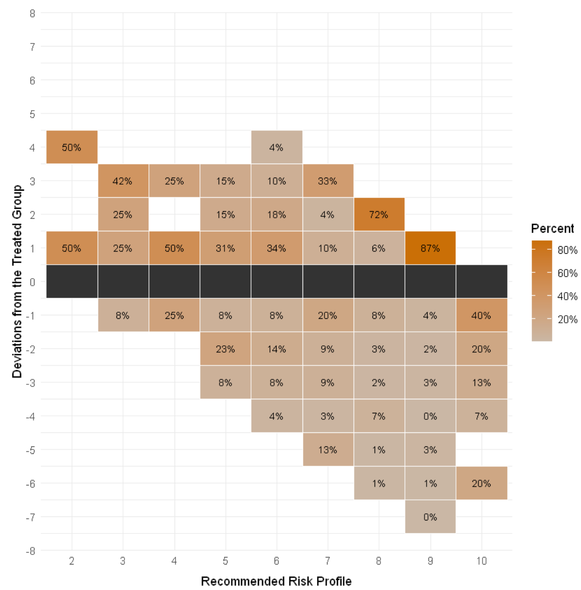
Figure 4 **Acceptance Rate per Recommended Risk Profile and Experimental Condition**

Descriptive statistics indicate that the overall acceptance rate does not differ significantly between the two experimental conditions: the control group exhibits a 54.63% acceptance rate, while the treatment group shows a 55.26% acceptance rate (two-sided Mann–Whitney U test, p = 0.666). Figure 4 displays the mean acceptance rate by risk profile for each group, revealing that both groups follow an almost identical pattern: acceptance rates range between approximately 80% and 90%

for risk profiles 2 through 6 and profile 10, decline to below 70% for profile 7, drop below 50% for profile 8, and fall below 40% for profile 9. Notably, risk profiles 7–9 account for roughly 72% of our observations (see Table 7 in Appendix B), underscoring the practical importance of these lower acceptance segments. Although previous XAI literature applied to robo-advisors predicts that providing explanations enhances the acceptance of algorithmic recommendations, our findings, both in aggregate and at the level of individual risk profiles, do not support this hypothesis.



(a) Deviations' Heatmap: Control Group



(b) Deviations' Heatmap: Treated Group

From the 2,093 observations in which users chose a different risk profile than recommended (1,645 positive deviations and 448 negative deviations), descriptive statistics reveal broadly similar deviation patterns across the two experimental conditions. The two heatmaps represent, for each recommended risk level, the percentage of users whose chosen profile lies at each deviation from the control group (Figure 5a) and the treated group (Figure 5b). In both graphs, deviations of +1 and +2 (i.e., selecting a risk profile one or two profiles higher than initially recommended) dominate at mid-range recommendations (profiles 3–7). For example, in the non-treated group, 42 percent of users recommended profile 3 actually selected profile 4 (+1 deviation), while 49 percent of users recommended profile 6 opted for profile 7 (+1). Similarly, in the treated group, 42 percent of profile 3 users and 33 percent of profile 6 users deviated upward by one notch. Upward deviations peak at recommended profile 8: 74 percent of non-treated users and 72 percent of treated users who deviated from it selected profile 10 (+2 deviation from 8). Notably, a substantial fraction of users recommended profiles 7 and 9 also deviated all the way to profile 10, suggesting they may have already had a predetermined preference for a fully equity-based portfolio. Negative deviations (selecting a lower risk profile than initially recommended) are relatively rare for mid-range recommendations but become more prevalent at the highest risk levels. For instance, among users recommended profile 10, 38 percent of treated users and 29 percent of non-treated users who deviated chose profile 4 (–6 deviation from 10), reflecting a strong downward shift for the riskiest recommendation. These patterns mirror the summary statistics in Table 8: both groups exhibit very similar weighted mean deviations for several recommended profiles (e.g., Profile 6 shows means of 0.367 in the non-treated group versus 0.400 in the treated group, and Profile 8 shows 0.907 versus 0.876), with modes typically at +1 for mid-range profiles. In other words, explanations did not substantially alter the direction or concentration of deviations once users elected to stray from the robo-advisor's suggestion.

4.2. Treatment Effect

Table 2 reports the results of logistic regressions on the likelihood of users accepting the Robo-Advisor's recommended risk profile ($Accept_i$), pooling all experimental conditions (Models 1–4). In each specification, we include a comprehensive set of socio-demographic and investment-related controls (as detailed in the note). Model 1 shows that the treatment indicator (i.e., provision of an explanation) has a small, positive coefficient (β = 0.014), but is not statistically significant (SE = 0.073). Across Models 2–4, we introduce interaction terms between the treatment indicator and two user characteristics (detailed in Section 4.1): device session (Phone) and user type (First-Timers). In Model 2, the interaction between Treatment and Phone is negative (β = –0.069) but statistically insignificant (SE = 0.148), leaving the marginal effect of explanations on phone users negligible. Model 3, First-Timers exhibit a significantly higher propensity to accept the recommendation (β = 0.327, SE = 0.119, $p < 0.01$), yet the main effect of explanations remains insignificant. In Model 4, which adds the Treatment × First-Timers interaction, the main effect of First-Timers decreases in magnitude and in significance (β = 0.259, SE = 0.141, $p < 0.10$) while the interaction term itself is statistically indistinguishable from zero (β = 0.132, SE = 0.151). Throughout all specifications, the Treatment coefficient does not attain significance, indicating that providing explanations does not meaningfully influence acceptance rates among users interacting with the same robo-advisor. By contrast, device session has a robust positive association with acceptance: users on mobile phones have approximately 37% higher odds to comply with the recommendation than users on a computer (Model 1: β = 0.314, SE = 0.078, $p < 0.01$), an effect that persists

Table 2 Explanations Effect on the Acceptance Rate

	Dependent variable:			
	$Accept_i$			
	(1)	(2)	(3)	(4)
Treated	0.014	0.049	0.016	−0.068
	(0.073)	(0.106)	(0.073)	(0.118)
Phone	0.314***	0.348***	0.317***	0.319***
	(0.078)	(0.108)	(0.078)	(0.078)
Treated × Phone		−0.069		
		(0.148)		
First-Timers			0.327***	0.259*
			(0.119)	(0.141)
Treated × First-Timers				0.139
				(0.151)
Constant	2.458**	2.437**	2.119*	2.161**
	(1.080)	(1.081)	(1.090)	(1.090)
Controls	Yes	Yes	Yes	Yes
Observations	4,645	4,645	4,645	4,645
Log Likelihood	−2,567.506	−2,567.380	−2,562.848	−2,562.366
Pseudo R^2 (McFadden)	0.197	0.197	0.198	0.198

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age; type of investment project; investment horizon; initially declared investment amount; rank of tunnel run; estimated revenue; home-ownership status; financial experience; financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

across all specifications. These findings suggest that, conditional on observables, neither device session nor user type moderates the null effect of explanations on acceptance. Mobile usage exerts an independent positive influence on compliance, while personal computer usage decreases it.

Table 3 presents OLS results for deviations (D_i) using again the 2,093 cases where users departed from the recommended profile. All four models control for demographics, investment characteristics, and initially recommended risk. Across all specifications, the treatment coefficient is small and not statistically different from zero. Likewise, being on a mobile device or a first-time user (and their interactions with treatment) shows no reliable impact on whether users shift toward higher or lower risk portfolios. The positive constant sign reflects the overall tendency to increase risk (1,645 positive vs. 448 negative deviations), but findings show that providing explanations does not shift that underlying direction.

Table 3 Explanations Effect on the Deviations

	Dependent variable:			
	D_i			
	(1)	(2)	(3)	(4)
Treated	0.040	−0.029	0.039	0.093
	(0.078)	(0.106)	(0.078)	(0.121)
Phone	−0.028	−0.098	−0.028	−0.030
	(0.079)	(0.111)	(0.079)	(0.079)
Treated × Phone		0.144		
		(0.156)		
First-Timers			−0.028	0.020
			(0.117)	(0.138)
Treated × First-Timers				−0.100
				(0.157)
Constant	0.702	0.757	0.733	0.693
	(1.099)	(1.107)	(1.102)	(1.098)
Controls	Yes	Yes	Yes	Yes
Observations	2,093	2,093	2,093	2,093
Adjusted R^2	0.221	0.221	0.221	0.221

Note: Robust standard errors are reported in parentheses. The regressions control for user age; type of investment project; investment horizon; initially declared investment amount; rank of tunnel run; estimated revenue; home-ownership status; financial experience; financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 9 in Appendix B reports OLS estimates for positive deviation (PD_i), including only the 1,645 cases where users chose a higher risk profile than recommended. All four regressions control for the same socio-demographic, investment variables, and initially recommended risk. In Model 1, the explanation treatment has a small, negative coefficient (β = –0.022, SE = 0.018) that is not statistically significant, indicating no appreciable shift in upward deviations. Adding the Treatment × Phone interaction in Model 2 does not change this result (β = –0.024, SE = 0.027; interaction β = 0.003, SE = 0.040). Likewise, Model 3 and 4 also shows that treated users exhibit a slightly smaller upward deviation (3, β = -0.022, SE = 0.018; 4, β = -0.024, SE = 0.027) but without significance, and Model 4's Treatment × First-Timers interaction is effectively zero (β = 0.003, SE = 0.038). Across all specifications, no coefficient on treatment, device, or user type meaningfully alters the magnitude of upward shifts.

Table 4 focuses on the 448 instances where users moved to a lower-risk portfolio than recommended, estimating how explanations affect the magnitude of these downward shifts. In Model

Table 4 Explanations Effect on the Negative Deviations

	Dependent variable:			
	ND_i			
	(1)	(2)	(3)	(4)
Treated	−0.282*	−0.452**	−0.176	0.070
	(0.147)	(0.189)	(0.146)	(0.240)
Phone	−0.068	−0.257	−0.053	−0.076
	(0.161)	(0.231)	(0.159)	(0.158)
Treated × Phone		0.390		
		(0.326)		
First-Timers			0.831***	1.044***
			(0.208)	(0.257)
Treated × First-Timers				−0.453
				(0.304)
Constant	−3.615***	−3.620***	−4.608***	−4.915***
	(0.643)	(0.640)	(0.687)	(0.716)
Controls	Yes	Yes	Yes	Yes
Observations	448	448	448	448
Adjusted R^2	0.134	0.136	0.156	0.158

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age; type of investment project; investment horizon; initially declared investment amount; rank of tunnel run; estimated revenue; home-ownership status; financial experience; financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

1, the coefficient on Treated is –0.282 (SE = 0.147, $p < 0.10$), indicating that, on average, users exposed to an explanation deviate below the prescribed risk level than those in the control group by 0.282 profile notches. In Model 2, once we introduce the Treated × Phone interaction, the main effect of Treated becomes even stronger (β = –0.452, SE = 0.189), suggesting that desktop users who see explanations reduce their chosen risk level by roughly half a notch more. The positive (albeit non-significant) interaction term (β = 0.390, SE = 0.326) implies that this "extra downward drift" vanishes for phone users, meaning the explanation-induced shift toward lower risk is concentrated among those on non-mobile devices. Turning to user types, Model 3 shows that First-Timers are, on their own, far less prone to drop so far below the recommended profile (β = 0.831, SE = 0.208, $p < 0.01$), effectively pulling their negative deviations closer to zero. However, the main treatment coefficient falls to –0.176 and becomes non-significant. When we add the Treated × First-Timers interaction in Model 4, the direct effect of Treated is also statistically indistinguishable from zero (β = 0.070, SE = 0.240), and the interaction (β = –0.453, SE = 0.304) remains insignificant, suggesting that explanations do not alter downward deviations among first-time users. In sum, explanations significantly magnify downward shifts in chosen risk, but primarily for returning users and computer users, while first-timers and phone users appear not to be influenced by the treatment.

In summary, transparency does not increase acceptance and does not change upward moves or overall engagement. The key margin is the device. Desktop sessions are less likely to follow the recommendation and deviate more from it than phone sessions. Conditional on that baseline, explanations amplify the downward move among desktop users who already choose a safer-than-recommended portfolio—i.e., treated desktop users shift even further toward safety relative to desktop controls. By contrast, phone sessions and first-time users show no detectable treatment response. Taken together, the evidence points to a clear narrative: compliance is higher on mobile; on desktop, users are more inclined to adjust—and the explanation nudges those downward-inclined users even farther down. Additionally, the overall deviation patterns, particularly the large shifts among those moving to a full-equity portfolio, suggest that many users may have already held a predetermined investment preference, rendering the robo-advisor's recommendation largely irrelevant regardless of treatment.

4.3. Additional analyses

Engagement outcomes Across all three engagement measures—total attempts (Table 10, Appendix B), "open" clicks to access the modification page (Table 11, Appendix B), and actual modification events (Table 12, Appendix B)—providing a peer-based graphical explanation had no impact on user behavior. In models predicting the total number of attempts, users who received an explanation neither reentered the tunnel process more nor fewer times than those in the control group, suggesting that explanations did not alter users' persistence or hesitancy during the onboarding process. Similarly, when examining how often participants clicked on the option to open the modification page, the only hint of a treatment effect was a small increase among treated desktop users when compared to untreated desktop users; however, this increase does not remain statistically significant once mobile usage and other covariates were taken into account, indicating no robust difference in exploratory behavior attributable to explanations. Finally, the count of actual modification events, instances where users applied changes to their risk selections, remained unchanged, implying that explanations did not spur users to explore other profiles by revising their choices more frequently. Engagement nonetheless differs systematically by device: phone sessions exhibit fewer openings of the modification option and fewer recorded modifications than desktop sessions (Tables 11–12; baseline coefficients around -0.34 for opens and -0.37 for modifications). This device pattern aligns with our main behavioral results: mobile users are more likely to accept the initial recommendation with minimal exploration, whereas desktop users explore more and, among those who move to a safer profile, explanations further increase the size of that downward deviation. Collectively, these findings demonstrate that, contrary to expectations from human–computer interaction and XAI research, displaying how the Robo-Advisor positions each user relative to the "average" risk profile neither increased nor decreased engagement. In other words, revealing prototypical characteristics for each risk class did not lead participants to explore more options, revisit the onboarding process more often, or investigate alternate portfolios. This suggests that engagement on this platform is driven by factors beyond algorithmic transparency, such as the type of device used.

Treatment heterogeneity Across both outcomes—acceptance and signed deviation—flexible CATE estimators confirm a near-zero average effect with only modest tails. Methodologically, we follow the framework in Jacob (2021): (i) Causal Random Forests grown with "honest" sample-splitting (one subsample to form leaves, a held-out subsample to estimate within-leaf treatment and control means), and (ii) Causal BART, which models the response surface as a sum of weak trees under regularizing priors and uses MCMC to obtain posterior draws of individual treatment effects. In our randomized setting, identification is straightforward (unconfoundedness holds by design); we nonetheless use honest splitting/held-out estimation to limit adaptive overfitting and obtain uncertainty via out-of-bag variance estimates for the forest and posterior credible intervals for BART. Tuning (number/depth of trees, priors, burn-in) follows Jacob (2021).

For acceptance, both methods yield tightly centered CATE distributions with ATEs around 0 with small positive and negative fringes (Figures 12a-12b; Table 13 in Appendix B). For deviation, both estimators produce CATE distributions that are still tightly centered near zero, with only modestly wider tails than for acceptance (Figures 13a–13b; Table 15 in Appendix B). The only consistent moderator is project objective: savings goals align with slightly more upward moves, retirement goals with slightly more downward moves (Table 16, Appendix B). By contrast, device (phone vs. computer), user tenure, age, baseline risk, and other demographics show no robust, method-consistent differences across CATE groups.

5. Conclusion

This paper uses a randomized controlled trial embedded in the interface of a leading French robo-advisor to test whether peer-based graphical explanations influence portfolio choices. Two facts emerge. First, explanations neither raise compliance with the recommended risk score nor do they increase engagement (tunnel completions, openings, or modifications). Second, device context impacts how explanations operate: relative to phone sessions, desktop sessions deviate more from the recommendation and explore more, and among desktop users who already move downward, explanations amplify that downward deviation.

These results carry implications for theory and practice. In high-stakes financial choices, where many investors appear to hold ex-ante portfolio preferences, algorithmic explanations may be too weak to overturn convictions or mitigate "algorithm aversion". They also point to limits of peer-based benchmarks: without interactive, personalized, or narrative components, explanations may fail to resonate with end-users. From a policy perspective, transparency mandates should not be assumed to yield greater adoption or trust absent complementary design choices.

On desktop, explanations appear to prompt more deliberation and caution, such that users who were already inclined to move to a safer profile choose even lower risk; on mobile—where adherence is higher and adjustments are rarer—explanations are largely inert. The takeaway is not that "explanations don't work", but that explanations work through context and can backfire precisely where users are most inclined to change course.

For design and policy, the implication is clear: transparency should be context-contingent, not one-size-fits-all. Platforms aiming to preserve adherence at the acceptance step should consider lighter or deferred explanations on desktop (e.g., progressive disclosure, post-confirmation details, or optional "learn more") and test benchmark framing that dampens conservative drift. Conversely, mobile flows may not warrant additional surface area for explanations at the decision point. More broadly, transparency alone may shift the margin of adjustment rather than the decision to accept.

Our setting bounds inference. We study a white-box, peer-based explanation during onboarding for life-insurance wrappers, and observe short-run choices rather than long-run funding, rebalancing, or performance. Other explanation styles (e.g., counterfactuals, conversational agents) or products may behave differently, and longer horizons may reveal delayed trust effects.

Overall, the evidence maps the boundary conditions of XAI in digital wealth: no average improvement in acceptance, higher adherence on mobile, and a desktop-specific backfire that pushes already-cautious users further down the risk ladder. Effective explainability in robo-advice will likely come from adaptive, device-aware designs—brief and unobtrusive when adherence is desired; richer and interactive when exploration is the goal—rather than uniform displays of feature importance. Future work should test such adaptive frameworks across investor segments and over extended horizons.

Notes

[1]Regulation (EU) 2016/679 (GDPR).

[2]The industry partner's identity is anonymized under an NDA that was a precondition for data access. Owing to confidentiality constraints and the partner's deployment timeline, the study was not preregistered. The intervention and its implementation were co-designed with the partner to ensure product feasibility; however, the authors retained full independence over the empirical analysis and the write-up. The partner had no editorial control and no right to approve or veto publication.

[3]For confidentiality reasons, the total number of users who entered the Tunnel and the exact start and end dates of the experiment cannot be disclosed. Consequently, our analysis focuses on those who signed a contract and is conducted at the contract level.

Appendix A: Experimental Design Appendices

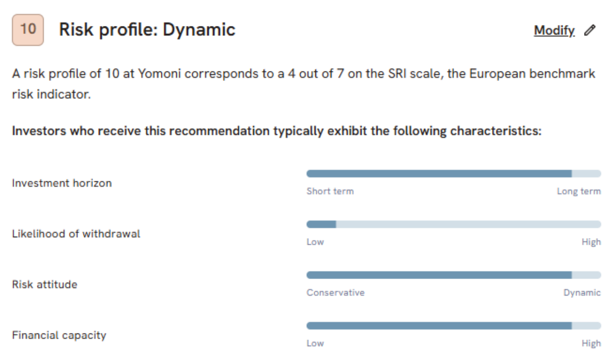


Figure 6 **Graphical explanation for the Risk Profile 10**

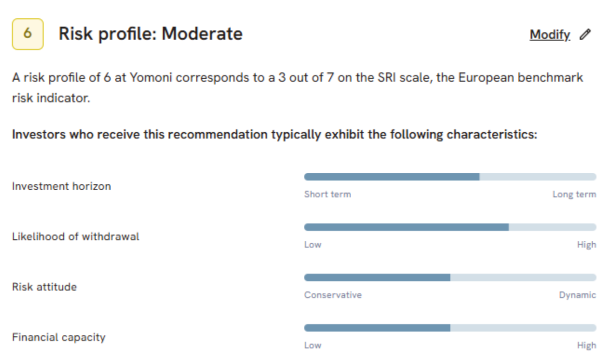


Figure 7 **Graphical explanation for the Risk Profile 6**

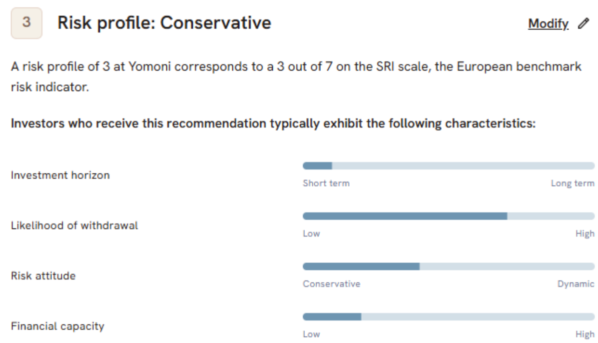


Figure 8 **Graphical explanation for the Risk Profile 3**

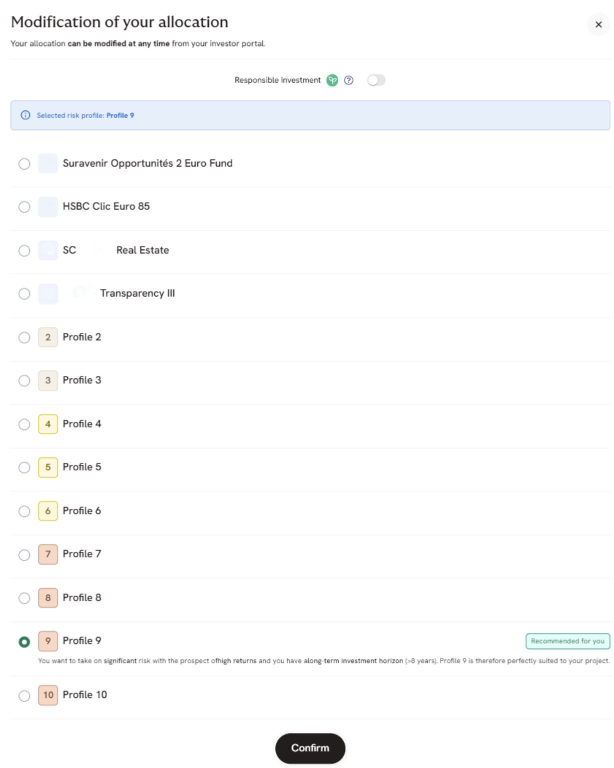


Figure 9 **Modification Page**

Appendix B: Additional Results

Acknowledgments

We would like to express our sincere gratitude to the participants of the University of Orléans Ph.D. Seminar and of the Society for Experimental Finance conference, for their comments and advice.

References

Abraham F, Schmukler S, Tessada J (2019) Robo-advisors: Investing through machines. Technical report.

Table 5 Balance tests between treatment and control groups (non-parametric)

Variable	p**-value**	**Bonferroni** p
Numeric / ordinal covariates (Mann–Whitney)		
User age	0.119	0.476
Initially declared investment amount	0.600	1.000
Rank of tunnel run	0.853	0.968
Initially recommended risk profile	0.749	1.000
Categorical covariates (χ^2 with Fisher fallback)		
Financial experience	0.625	1.000
Estimated Revenue	0.767	1.000
Type of investment project	0.571	1.000
Home-ownership status	0.872	1.000
Financial literacy	0.978	1.000
Investment Horizon	0.761	1.000

Notes: Two-sided tests. Numeric/ordinal covariates tested with Mann–Whitney (Wilcoxon rank-sum). Categorical covariates tested with Pearson χ^2; Fisher's exact used when expected counts are small. Bonferroni adjustment applied within each family (numeric vs. categorical). Values rounded to three decimals.

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access* 6:52138–52160.

Anjomshoae S, Najjar A, Calvaresi D, Främling K (2019) Explainable agents and robots: Results from a systematic literature review. *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, 1078–1088 (International Foundation for Autonomous Agents and Multiagent Systems).

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. (2020) Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion* 58:82–115.

Beaudouin V, Bloch I, Bounie D, Clémençon S, d'Alché Buc F, Eagan J, Maxwell W, Mozharovskyi P, Parekh J (2020) Flexible and context-specific ai explainability: a multidisciplinary approach. *arXiv preprint arXiv:2003.07703* .

Ben David D, Resheff YS, Tron T (2021) Explainable ai and adoption of financial algorithmic advisors: an experimental study. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 390–400.

Bianchi M, Brière M (2020) Robo-advising for small investors. *SSRN Electronic Journal* .

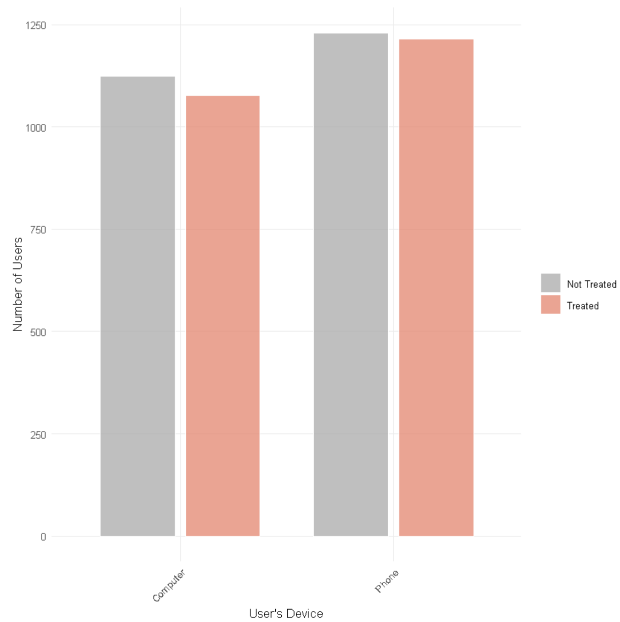


Figure 10 **Distribution of Devices Used by Experimental Condition**

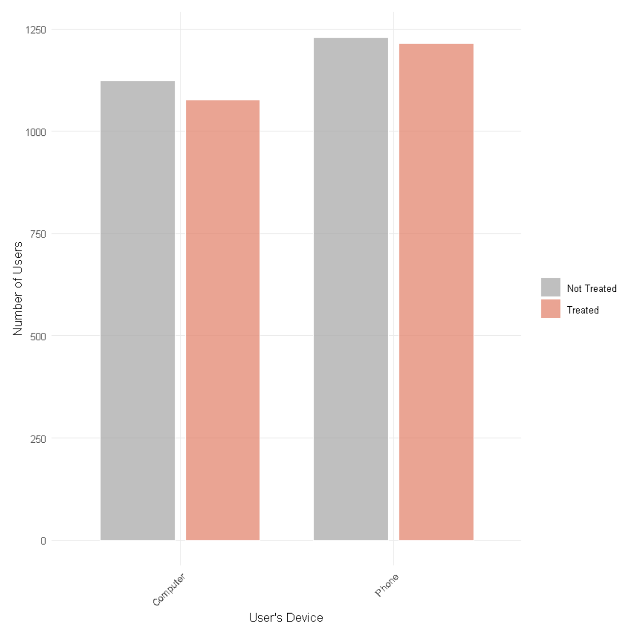


Figure 11 **Distribution of Users' Type by Experimental Condition**

Bianchi M, Briere M (2021) Robo-advising: less ai and more xai? *Available at SSRN 3825110* .

Board FFS (2017) *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications* (Financial Stability Board).

Cai CJ, Jongejan J, Holbrook J (2019) The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th international conference on intelligent user interfaces*, 258–262.

Table 6 Summary Statistics for Behavioral and Engagement Outcomes

Outcome	Q1	Mean	Median	Q3	SD	N
Behavioral Outcomes						
Acceptance ($Accept_i$)	0.00	0.55	1.00	1.00	0.49	4,645
Deviation (D_i)	1.00	0.52	1.00	2.00	1.91	2,093
Positive deviation (PD_i)	1.00	1.39	1.00	2.00	0.64	1,645
Negative deviation (ND_i)	-4.00	-2.65	-2.00	-1.00	19.04	448
Engagement Outcomes						
Total attempts ($TotAttempts_i$)	1.00	1.48	1.00	2.00	0.99	3,856
Opening events ($OpenEvents_i$)	0.00	1.14	0.00	1.00	3.69	3,856
Modification events ($ModEvents_i$)	0.00	0.71	0.00	0.00	2.83	3,856

Table 7 Number of observations per recommended risk level by experimental condition and in total.

Risk Profile	Not Treated	Treated	Total	Overall (%)
Profile 2	51	42	93	2.0
Profile 3	121	119	240	5.2
Profile 4	51	34	85	1.8
Profile 5	67	87	154	3.3
Profile 6	245	252	497	10.7
Profile 7	191	200	391	8.4
Profile 8	582	538	1120	24.1
Profile 9	943	926	1869	40.2
Profile 10	103	93	196	4.2

Capponi A, Olafsson S, Zariphopoulou T (2022) Personalized robo-advising: Enhancing investment through client interaction. *Management Science* 68(4):2485–2512.

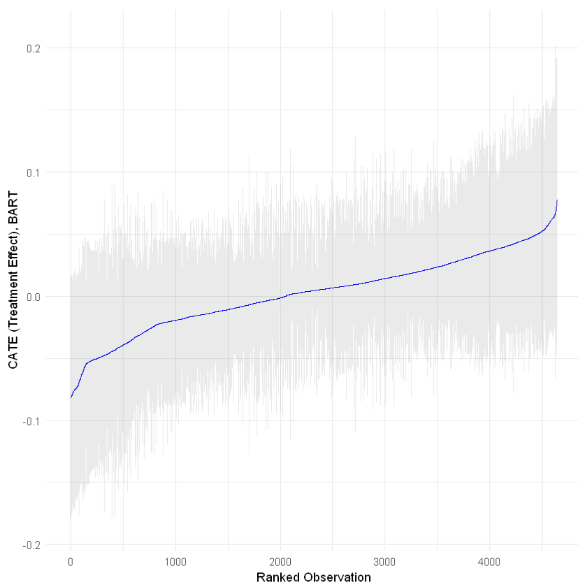
Castelvecchi D (2016) Can we open the black box of ai? *Nature* 538(7623):20–23.

Cheng HF, Wang R, Zhang Z, O'connell F, Gray T, Harper FM, Zhu H (2019) Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–12.

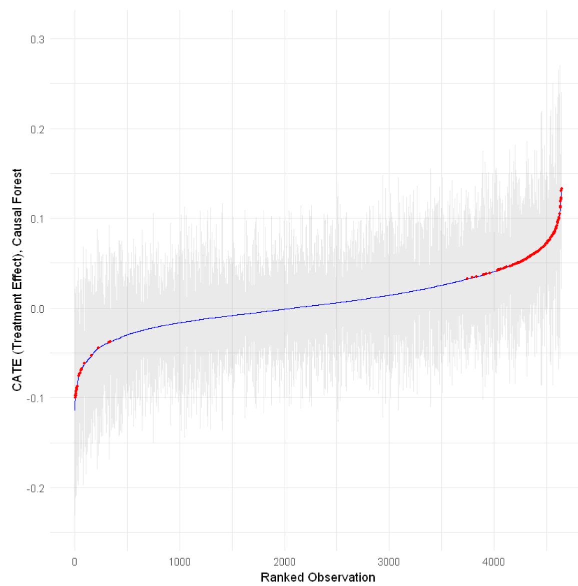
Cohen SB, Dror G, Ruppin E (2005) Feature selection based on the shapley value. *Proceedings of IJCAI*, 1–6.

Table 8 Summary Statistics of Deviation by Recommended Risk Level and Experimental Condition

	Non-treated			Treated		
Risk Profile	**Total** n	**Weighted Mean**	**Mode**	**Total** n	**Weighted Mean**	**Mode**
Profile 2	7	2.7143	1	4	2.5000	1
Profile 3	12	2.0833	3	12	1.9167	3
Profile 4	9	0.0000	1	4	1.0000	1
Profile 5	5	–0.2000	–2	13	0.3077	1
Profile 6	46	0.3696	1	50	0.4000	1
Profile 7	67	0.8507	3	70	–0.2143	3
Profile 8	300	0.9067	2	275	0.8764	2
Profile 9	601	0.4093	1	582	0.5206	1
Profile 10	21	–3.8095	–6	15	–2.6667	–1



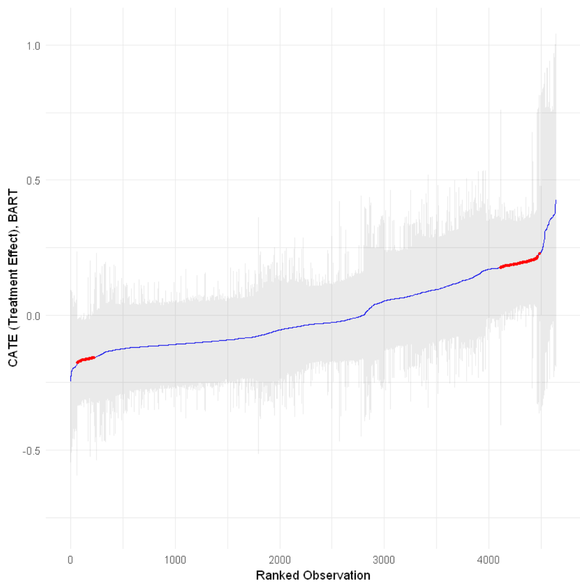
(a) Causal BART, Acceptance of the Recommendation



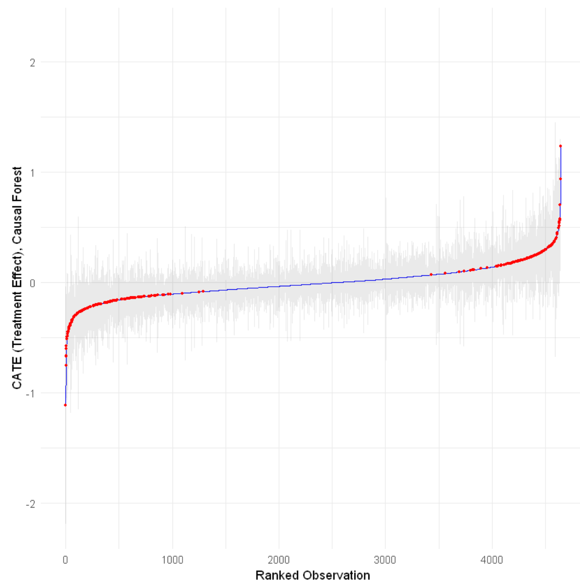
(b) Causal Forest, Acceptance of the Recommendation

Covert I, Lundberg S, Lee SI (2021) Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research* 22(209):1–90.

Cramer H, Evers V, Ramlal S, Van Someren M, Rutledge L, Stash N, Aroyo L, Wielinga B (2008) The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction* 18:455–496.



(a) Causal BART, Deviation



(b) Causal Forest, Deviation

D'acunto F, Rossi A (2022) Fintech and robo-advising: The transformation role of ai in personal finance. *SSRN Electronic Journal* URL `http://dx.doi.org/10.2139/ssrn.4100784`.

Deo S, Sontakke NS (2021) Usability, user comprehension, and perceptions of explanations for complex decision support systems in finance: A robo-advisory use case. *Computer* 54(10):38–48.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144(1):114.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science* 64(3):1155–1170.

Eslami M, Krishna Kumaran SR, Sandvig C, Karahalios K (2018) Communicating algorithmic process in online behavioral advertising. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–13.

European Commission (2019) Ethics guidelines for trustworthy AI. URL `https://ec.europa.eu/futurium/en/ai-alliance-consultation`, available at: `https://ec.europa.eu/futurium/en/ai-alliance-consultation`.

Fein ML (2017) Are robo-advisors fiduciaries? *Available at SSRN 3028268* .

Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 241–250.

Hong CY, Lu X, Pan J (2020) Fintech adoption and household risk-taking: From digital payments to platform investments. Technical report, National Bureau of Economic Research.

Hu D, Guo F, Shang J, Zhang X (2024) Does digital finance increase household risk-taking? evidence from china. *International Review of Economics & Finance* 93:1197–1210.

Jacob D (2021) Cate meets ml: Conditional average treatment effect and machine learning. *Digital Finance* 3(2):99–148.

Jung D, Dorner V, Weinhardt C, Pusmaz H (2018) Designing a robo-advisor for risk-averse, low-budget consumers. *Electronic Markets* 28(3):367–380, URL `http://dx.doi.org/10.1007/s12525-017-0279-9`.

Kahneman D (2013) A perspective on judgment and choice: Mapping bounded rationality. *Progress in Psychological Science around the World. Volume 1 Neural, Cognitive and Developmental Issues.* 1–47.

Krzyziński M, Spytek M, Baniecki H, Biecek P (2023) Survshap (t): time-dependent explanations of machine learning survival models. *Knowledge-Based Systems* 262:110234.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Lurie NH, Berger J, Chen Z, Li B, Liu H, Mason CH, Muir DM, Packard G, Pancras J, Schlosser AE, et al. (2018) Everywhere and at all times: mobility, consumer decision-making, and choice. *Customer Needs and Solutions* 5:15–27.

Mograbi E (2022) Decision-makers are more impulsive on smartphones than on computers. *Journal of Behavioral and Experimental Economics* 100:101916.

Morana S, Gnewuch U, Jung D, Granig C (2020) The effect of anthropomorphism on investment decision-making with robo-advisor chatbots. *ECIS*.

Nunes I, Jannach D (2017) A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27:393–444.

Patel K, Lincoln M (2019) It's not magic: Weighing the risks of ai in financial services. *Report available at https://www. european-microfinance. org/publication/its-not-magic-weighing-risks-ai-financial-services* .

Philippon T (2019) On fintech and financial inclusion. *National Bureau of Economic Research* URL `http://dx.doi.org/10.3386/w26330`.

Preece A, Harborne D, Braines D, Tomsett R, Chakraborty S (2018) Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184* .

Rader E, Cotter K, Cho J (2018) Explanations as mechanisms for supporting algorithmic transparency. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–13.

Ribeiro MT, Singh S, Guestrin C (2016) "why should I trust you?": Explaining the predictions of any classifier. *CoRR* abs/1602.04938, URL `http://arxiv.org/abs/1602.04938`.

Riedl MO (2019) Human-centered artificial intelligence and machine learning. *Human behavior and emerging technologies* 1(1):33–36.

Shen Y, Hu W, Zhang Y (2022) Digital finance, household income and household risky financial asset investment. *Procedia Computer Science* 202:244–251.

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies* 146:102551.

Speith T (2022) A review of taxonomies of explainable artificial intelligence (xai) methods. *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2239–2250.

Strzelczyk BE (2017) Rise of the machines: The legal implications for investor protection with the rise of robo-advisors. *DePaul Bus. & Comm. LJ* 16:54.

Wang RJH, Malthouse EC, Krishnamurthi L (2015) On the go: How mobile shopping affects customer purchase behavior. *Journal of retailing* 91(2):217–234.

Weitz K, Schiller D, Schlagowski R, Huber T, André E (2021) "let me explain!": exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces* 15(2):87–98.

Table 9 Explanations Effect on the Positive Deviations

	Dependent variable:			
	PD_i			
	(1)	(2)	(3)	(4)
Treated	−0.022	−0.024	−0.022	−0.024
	(0.018)	(0.027)	(0.018)	(0.027)
Phone	0.007	0.005	0.007	0.007
	(0.019)	(0.031)	(0.019)	(0.019)
Treated × Phone		0.003		
		(0.040)		
First-Timers			−0.007	−0.008
			(0.027)	(0.035)
Treated × First-Timers				0.003
				(0.038)
Constant	2.136*	2.137*	2.143*	2.144*
	(1.174)	(1.176)	(1.175)	(1.171)
Controls	Yes	Yes	Yes	Yes
Observations	1,645	1,645	1,645	1,645
Adjusted R^2	0.668	0.668	0.668	0.668

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age; type of investment project; investment horizon; initially declared investment amount; rank of tunnel run; estimated revenue; home-ownership status; financial experience; financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 10 Explanations Effect on the Total Number of Attempts

	Dependent variable:			
	$TotAttempts_i$			
	(1)	(2)	(3)	(4)
Treated	−0.007	0.009	−0.001	0.002
	(0.022)	(0.030)	(0.020)	(0.034)
Phone	0.012	0.027	0.013	0.013
	(0.023)	(0.031)	(0.021)	(0.021)
Treated × Phone		−0.029		
		(0.042)		
First-Timers			0.726***	0.728***
			(0.031)	(0.038)
Treated × First-Timers				−0.004
				(0.041)
Constant	0.149	0.141	−0.577***	−0.578***
	(0.143)	(0.145)	(0.137)	(0.137)
Controls	Yes	Yes	Yes	Yes
Observations	3,856	3,856	3,856	3,856
Log Likelihood	−5,042.480	−5,042.332	−4,912.462	−4,912.459
Pseudo-R^2 (McFadden)	0.030	0.030	0.055	0.055

Note: Robust standard errors are reported in parentheses. The regressions control for user age; type of investment project; investment horizon; initially declared investment amount; rank of tunnel run; estimated revenue; home-ownership status; financial experience; financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 11 Explanations Effect on the Total Number of Open Events

	Dependent variable:			
	$OpenEvents_i$			
	(1)	(2)	(3)	(4)
Treated	0.065	0.212*	0.071	0.102
	(0.101)	(0.123)	(0.101)	(0.171)
Phone	−0.340***	−0.171	−0.337***	−0.338***
	(0.094)	(0.139)	(0.093)	(0.092)
Treated × Phone		−0.336*		
		(0.195)		
First-Timers			0.839***	0.867***
			(0.139)	(0.177)
Treated × First-Timers				−0.056
				(0.209)
Constant	−1.084*	−1.170*	−1.970***	−1.989***
	(0.609)	(0.609)	(0.658)	(0.664)
Controls	Yes	Yes	Yes	Yes
Observations	3,856	3,856	3,856	3,856
Log Likelihood	−7,918.582	−7,904.173	−7,779.042	−7,778.649
Pseudo-R^2	0.141	0.142	0.156	0.156
Note:				*p<0.1; **p<0.05; ***p<0.01

Note: Robust standard errors are reported in parentheses. The regressions control for user age; type of investment project; investment horizon; initially declared investment amount; rank of tunnel run; estimated revenue; home-ownership status; financial experience; financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 12 Explanations Effect on the Total Number of Modification Events

	Dependent variable:			
	$ModEvents_i$			
	(1)	(2)	(3)	(4)
Treated	0.051	0.202	0.056	0.171
	(0.117)	(0.142)	(0.117)	(0.189)
Phone	-0.366^{***}	-0.191	-0.362^{***}	-0.366^{***}
	(0.108)	(0.163)	(0.108)	(0.107)
Treated × Phone		-0.349		
		(0.224)		
First-Timers			0.747^{***}	0.850^{***}
			(0.174)	(0.215)
Treated × First-Timers				-0.211
				(0.234)
Constant	-1.689^{**}	-1.777^{***}	-2.484^{***}	-2.550^{***}
	(0.680)	(0.683)	(0.728)	(0.734)
Controls	Yes	Yes	Yes	Yes
Observations	3,856	3,856	3,856	3,856
Log Likelihood	−5,676.116	−5,666.519	−5,607.795	−5,604.267
Pseudo-R^2	0.152	0.154	0.162	0.163

Note: Robust standard errors are reported in parentheses. The regressions control for user age; type of investment project; investment horizon; initially declared investment amount; rank of tunnel run; estimated revenue; home-ownership status; financial experience; financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 13 CATE results for the Acceptance of the Recommendation

Method	20% least	ATE	20% most
Causal BART	-0.0381	0.0023	0.0429
Causal Forest	-0.0368	0.0059	0.0532

Table 14 Characteristics of Users by CATE Group, Acceptance of the Recommendation

Variable	No Significance (n=4,502)	Positive (n=119)	Negative (n=24)	p-value
Proportions				
Low patrimony (<30 k)	0.1966	0.0924	0.2917	0.0087**
Project Type: Savings	0.8827	0.9664	0.9167	0.0164*

Note: Proportions are means of 0/1 indicators. p-values from one-way ANOVA across the three groups. *p<0.05; **p<0.01.

Table 15 CATE results for the Deviation

Method	20% least	ATE	20% most
Causal BART	-0.1368	0.0003	0.1979
Causal Forest	-0.1938	-0.005	0.2052

Table 16 User Characteristics by CATE Group and Method (Deviation)

	BART				Causal Forest			
Variable	NoSig	Pos	Neg	p(BART)	NoSig	Pos	Neg	p(CF)
n	4,347	210	88	—	4,454	53	138	—
Proportions								
Financial Knowledge: Beginner	0.445	0.409	0.329	0.063*	—	—	—	—
Financial Knowledge: Knowledgeable	0.393	0.433	0.500	0.068*	—	—	—	—
Investment Horizon	10.806	11.685	11.579	0.065*	—	—	—	—
Project type: Savings	0.883	0.933	0.863	0.068***	0.885	0.940	0.821	< 0.01***
Project type: Retirement	—	—	—	—	0.044	0.006	0.076	< 0.01***
Means								
User age (years)	—	—	—	—	21.54	17.90	23.22	< 0.01***
Recommended risk profile	6.615	6.985	6.909	< 0.01***	—	—	—	—

Note: "NoSig," "Pos," and "Neg" denote non-significant, positive, and negative CATE estimates, respectively; proportions for binary indicators, means for continuous variables. p-values from one-way ANOVA across the three CATE groups, by method. *p<0.10;**p<0.05; ***p<0.01.