Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making

Astrid Bertrand, James Eagan, Winston Maxwell

Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making

Astrid Bertrand
astrid.bertrand@telecom-paris.fr
LTCI, Institut Polytechnique de Paris
France

Winston Maxwell
winston.maxwell@telecom-paris.fr
i3, CNRS, Institut Polytechnique de Paris
France

James R. Eagan
james.eagan@telecom-paris.fr
LTCI, Institut Polytechnique de Paris
France

ABSTRACT

Robo-advisors are democratizing access to life-insurance by enabling fully online underwriting. In Europe, financial legislation requires that the reasons for recommending a life insurance plan be explained according to the characteristics of the client, in order to empower the client to make a "fully informed decision". In this study conducted in France, we seek to understand whether legal requirements for feature-based explanations actually help users in their decision-making. We conduct a qualitative study to characterize the explainability needs formulated by non-expert users and by regulators expert in customer protection. We then run a large-scale quantitative study using Robex, a simplified robo-advisor built using ecological interface design that delivers recommendations with explanations in different hybrid textual and visual formats: either "dialogic"—more textual—or "graphical"—more visual. We find that providing feature-based explanations does not improve appropriate reliance or understanding compared to not providing any explanation. In addition, dialogic explanations increase users' trust in the recommendations of the robo-advisor, sometimes to the users' detriment. This real-world scenario illustrates how XAI can address information asymmetry in complex areas such as finance. This work has implications for other critical, AI-based recommender systems, where the General Data Protection Regulation (GDPR) may require similar provisions for feature-based explanations.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

explainability, intelligibility, AI regulation, financial inclusion

ACM Reference Format:
Astrid Bertrand, Winston Maxwell, and James R. Eagan. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3593013.3594053

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594053

1 INTRODUCTION

As online AI-based services are becoming more ubiquitous with commercial recommender systems, internet users are exposed to opaque personalized suggestions. This raises questions on how to communicate relevant and accessible information to foster appropriate trust in those systems [8]. While explanations are often unnecessary or non-critical in many low-risk applications of AI, such as for movie or music suggestions, they can be mandated by law in some high-stakes industries, such as finance, through the legal notion of "informed decision".

Real-world scenarios of explainability in the scientific literature are primarily in the health care domain [9, 19, 20, 24]. In this paper, we focus on another use case of explainability which is equally high-stake, widespread, and legally motivated: AI-based financial advice, *i.e.* robo-advisors. Explanations of these systems are required to make online services to savings and investment customers more understandable. The challenge is to ensure that customers are informed of the processes by which a recommendation is made, through clear explanations. This aims at protecting clients from recommendations misaligned with their objectives, risk appetite and other personal characteristics. Moreover, the financial domain can feel overwhelming and complex to many people [38], which poses an additional challenge: explaining in simple terms not only the attributes of the system but also financial principles to novice users. Few studies [6] have focused on how to design legally mandated explanations for lay users in real-world, high-stakes scenarios. Yet, the lack of understanding of how explainability requirements should be implemented is currently a barrier to the use of AI systems in high stake domains [5]. We aim to address this gap by leveraging the knowledge of customer protection specialists about existent explainability requirements in the financial domain. We interviewed 6 customer protection experts who work at the French regulatory authority of financial services to describe the legal motivations and expectations for explanations in this domain and test the propensity of feature-based explanations to meet these requirements. We believe the insights from experts from the regulatory sphere present interesting yet so far unsolicited proxies for characterizing the users' needs. Our aim is to better understand the regulatory challenges arising with explainability, which we believe is an under-explored area in the human-computer interaction side of the XAI field. Our first research question is the following:

RQ1: What are the regulatory expectations for explanations in financial investment services to protect customers? How can current XAI methods meet them?

In addition, we interviewed 5 lay users on their needs for explanations of robo-advisors. This enabled us to qualitatively compare

regulatory and "practical" needs for explanations, in an attempt to address the second research question:

RQ2: How do regulators on the one hand and end users on the other describe the need for explanations?

To illustrate how legal requirements might be transformed into explanation representations, we designed several formats of feature importance explanations and conducted a large-scale study with 256 participants to compare their impact on user trust, and users' appropriate reliance and understanding. Recent advances in the fast-growing field of explainability have brought a better understanding of how different representations and interactions of AI explanations impact non-expert users [7, 10, 35, 39, 47]. Szymanski *et al.* [47] found that lay users preferred graphical explanations but could more easily misinterpret them compared to textual explanations, motivating the need for hybrid textual and visual explanations. However, little is known about where the cursor should be placed between textual and visual content. In this paper, we compare different formats of hybrid textual and graphical explanations using SHAP [30]. Our aim is to answer the following research question:

RQ3: How effective are different representations of hybrid textual and graphical explanations to protect non-expert users?

The contributions of this paper can be summarized as follows. We analyze the legal requirements for explainability in a real-world context: online life-insurance underwriting. Then, in a qualitative study, we compare regulators' and end-users' perspectives on legal explainability requirements in life-insurance and argue for the relevance of consulting regulators for defining customers' XAI needs. Finally, we provide evidence through a large-scale study that the benefits of explanations on user understanding, appropriate trust and reliance are not clear, and that dialogic explanations might lead to harmful over-reliance.

2 RELATED WORK

2.1 Understanding explainability needs

In recent years, the XAI community has made substantial progress in making AI systems more intelligible to end users [22, 27, 29, 42, 51]. Much of this work aimed at understanding user needs to better inform the design of technical solutions [15, 26, 27, 34]. Using semi-structured interviews, articles such as [27, 28, 45] give an account of users' questions and motivations regarding explainability. They inform on the actual user demand for information about AI systems by presenting taxonomies of user questions [27, 28], for example. Theoretical approaches have also provided important insights on users' cognitive needs regarding explainability in the form of frameworks, surveys or theories [34, 44, 46, 48, 51]. For example, Miller [34] draws on how humans explain things to each other to find out what people expect from explanations.

All these studies provide relevant findings to inform on the actual needs of users regarding explainability. Another potentially relevant source of information to design helpful explanations are legal requirements. Very few XAI research efforts have been motivated by legal obligations to produce explanations such as the "right to explanation" included in the GDPR. Bibal *et al.* [6] give a complete overview of existing legal frameworks for explainable AI. However, the point of view of regulators has not been solicited so far in the explainability literature, to the best of our knowledge.

2.2 Representing AI explanations to non-expert users

2.2.1 Explanation formats. A few contributions from the computer science side of XAI conducted user studies to evaluate the ability of XAI methods to successfully convey accurate mental models of AI systems to users. In particular, this line of research sheds light on the limitations of some technical solutions for aiding user understanding, or worse, on their potential for deception [21, 23, 40]. Some work has focused specifically on the implementation of explanations for non-expert users in specific contexts [7, 10, 47]. Cheng *et al.* [10] presented explanations of an algorithmic school admission decision process to users with no domain or technical expertise. They found that static and interactive explanations, where users could change the inputs to see the resulting outcome, improved users' understanding of the AI decisions. Bove *et al.* [7], however, were unable to replicate these results in the context of explaining an algorithmic car insurance pricing decision. They did not find that explanations improved comprehension but they did improve user satisfaction. Szymanski *et al.* [47] studied how different representations of explanations (either visual, textual or both) affect users' understanding of an AI system in an artificial task (estimating the reading time of news articles). The paper shows that purely visual explanations (in this case line graphs) can be subject to misinterpretation, while purely textual explanations are better understood but less satisfactory to users. A combination of the two representations could provide the best of both worlds. However, there could be many different ways to design "hybrid" textual and visual explanations. Specifically, it is still unclear if textual explanations presented as conversations achieve better user preferences and improve task accuracy compared to graphical formats.

Additionally, explanations' ability to engage users in a sensitive and complex topic such as financial investment has not yet been studied in the XAI literature where artificial contexts are often used as test benches [8, 11, 14].

2.2.2 Mitigating overreliance issues. Other work in human-centered XAI research has been studying how expertise affects the perception of explanations. For example, Simkute *et al.* [43] stress the importance of differentiating the reasoning of experts from that of lay users and reflecting this difference in the design of explanations. Quite logically, experts are able to be more critical of the explanations, sometimes at the cost of not trusting them enough, whereas lay users are more subject to over-reliance [3, 41]. Eiband *et al.* [12], for example, demonstrated that the mere presence of explanations reinforced non experts' trust using placebic explanations.

Explanations must therefore support either trust building for experts, or critical thinking for lay users. Another key difference is the level of motivation to use explanations, which can be much lower for non-expert users. This makes it particularly challenging to make explanations both simple and appealing to lay users, while encouraging cognitive engagement and skepticism [4, 36]. It is still unclear if explanations for non-expert users can be designed to foster trust and understanding on the one hand while encouraging users' critical thinking (*i.e.* ability to detect errors) on the other.

This might be desirable in sensitive contexts where the algorithmic output can have strong consequences on the user's life quality.

3 THE TEST-BED FOR STUDYING EXPLANATIONS OF RECOMMENDATIONS OF FINANCIAL CONTRACTS

In this paper, we focus on a real-case application of explainability: explanations of online recommendations for life insurance products. In Europe, explanations in this context are legally required by sector-specific regulations to ensure customer protection. We describe below the case study context, the related legal requirements for explanations and the system used in the studies presented in Section 4 and Section 5.

3.1 Context

3.1.1 Life-insurance underwriting. As AI systems gain performance, their adoption expands to areas considered critical. In finance, increasingly sophisticated recommender systems known as "robo-advisors" are democratizing online underwriting of life insurance. In France, where the study was conducted, life insurance is a savings vehicle used both to pass on money to a designated beneficiary upon the death of the subscriber of the contract, and to make a long-term financial investment in a tax-advantaged environment. In the rest of the paper, we will only address the latter, most common usage of life-insurance. Life insurance subscribers are presented with a financial recommendation with a specific level of risk (a higher level of risk means more chances to win big but also more chances to lose). Choosing a life insurance contract with an appropriate risk level—not too high for the client's financial situation—is crucial to ensuring clients' financial stability. However, many clients may not be financially literate. Therefore, French and European legislation[1] require insurance providers to produce "clear, precise and non-misleading" explanations to guide potential customers towards an "informed" decision and address the asymmetry of information between client and advisor. We describe further the legal requirements to explain recommendations in this context in the next section. Most existing online recommender systems currently fall short of this legal explanation requirement, according to our discussions with French regulators in the life-insurance sector. Specifically, explanations of online recommender systems, *i.e.* robo-advisors, rarely focus on the reasons why a recommendation is adapted to the user's need, which is the type of explanation we focus on in this paper.

3.1.2 Towards more digital and AI powered systems. The automated advice provided by robo-advisors is seen as a more cost-efficient way to deliver proposals to pockets of population who do not otherwise have access to financial advice, as an OECD report highlights [31]. Additionally, the COVID crisis has accelerated the interest in online systems with the increasing demand for online and real-time services [2]. As seen through our series of interviews with regulators in life-insurance—described later in the paper—, most current robo-advisors (specifically in France where this study was conducted) are rule-based, with varying degrees of complexity in the amount and nature of the rules. Yet, many studies foresee an acceleration of AI-based underwriting solutions in the financial sector and in life-insurance [2, 31]. AI-powered systems offer faster and more personalized financial advice. For brokers, data-driven underwriting helps identify risk in a more fine-grained manner [1]. The insurance market is also gaining interest in AI-powered robo-advisors with the successful examples of companies which used this technology to increase sales revenue significantly [1].

3.1.3 Legal Requirements for feature-based explanations. In the life-insurance context, financial legislation regarding the insurance sector apply. The law on insurance distribution (Articles 20 and 30 of Directive (EU) 2016/97 of January 20, 2016), which aims to protect consumers against the sale of products unsuited to their needs, requires providers to explain "the reasons for the appropriateness of the proposed contract". Our research question is which explanation format, especially provided by automatic means—through a roboadvisor—, is the most suitable format to protect the consumer. This leads us to question more precisely the purpose of the explanation in light of the objectives of the law. What exactly is expected of the explanation so that it is effective with regard to the objectives of the Articles L. 521-4 and L. 522-5 of the French Insurance Code and EU Directive 2016/97? One of the objectives of the explanations is to enable future life-insurance subscribers to make a "fully informed" decision about the product being proposed. This objective is explicitly stated in the text of Article L. 521-4 of the French Insurance Code and Article 20 of EU Directive 2016/97. However, this objective is relatively imprecise and difficult to measure. To better measure whether an explanation allows for an "informe" decision, the goal should be broken down into subgoals that are easier to test and measure. We understand these subgoals to be 1) help users appropriately rely on a recommendation (and be able to detect a big mistake) 2) help users understand the appropriateness of a recommendation for them 3) help users calibrate their trust in robo-advisors. This is therefore what we measured in Study 2.

In addition to the goal of "fully informing" clients, the law pursues the objective of supervising the behavior of intermediaries by imposing the obligation to set out in writing the client's needs as well as the reasons why the recommended product is in line with those needs. The formalization of these steps will reduce the risks of intermediaries taking shortcuts and letting conflicts of interest interfere with their duty to give objective investment advice to customers.

In other contexts, AI systems may also be affected by requirements for feature-based explanations. Consumer protection law has provisions regarding explanations of recommender systems in online marketplaces. It notably imposes to show "the main parameters determining the ranking [...] of offers presented to the consumer as a result of the search query and the relative importance of those parameters as opposed to other parameters"[2]. Moreover, the GDPR provisions can also apply in some contexts. It requires that data controllers disclose "meaningful information about the logic involved" (articles 13-15) in entirely automated decisions. The GDPR provisions apply "when the decisions (i) involve the processing of personal data, (ii) are based solely on an automated processing of

[1]The European Parliament and the European Concil. 2016. Directive (EU) 2016/97 on insurance distribution.

[2]New art. 6(a) of Directive 2011/83 on Consumer Rights

Figure 1: Fictional life-insurance plans proposed by Robex, the explainable robo-advisor developed for this study

data and (iii) produce legal or significant effects on the recipient of the decision" [6].

3.2 Robex, the explainable robo-advisor

3.2.1 A simplified model. Robex—standing for EXplainable ROBo-advisor—is a simplified and fictional life-insurance recommender system developed for the purpose of this study. Robex's recommendation algorithm is not AI but a rule-based algorithm established with the help of domain experts. Indeed, since our goal was to study explanation representations using existing agnostic explainability methods, we did not need to use a real AI algorithm for this study. The design of Robex was done using Ecological Interface Design [50]. We reviewed existing robo-advisors and conducted informal interviews with 4 regulators with extensive experience in the control of intermediaries (or brokers) in life-insurance to better understand the domain. Based on these discussions, we developed a profiling questionnaire to measure 5 user characteristics: the amount to be invested compared to the user's total financial wealth, her investment objective, her financial knowledge and experience, her risk appetite and the proportion of her financial assets already placed on financial markets. For each of the questions used to measure these characteristics (cf. Table 3 of the Appendix), we associated coefficients so as to obtain a risk-score that denoted the amount of risk a user can take. We were then able to sketch five fictional but realistic life-insurance plans that represented 5 levels of risk. Our score-based, simplified underwriting rules then matched a profile to a plan.

The usual underwriting process with robo-advisors—and Robex—is as follows. First, users go through a series of questions about their profile and financial objectives. Then, they can see the summary of their profile and the proposed recommendation—on the same page in Robex. During this recommendation phase, Robex presents an additional section on why this product is recommended to you.

3.2.2 Feature Importance Explanations. We approached the explainability phase as if the Robex algorithm was a black-box, so that our results can be transposed to more opaque AI-powered robo-advisors. As seen in Section 3.1.3, the required explanations in life-insurance but also for other online recommender systems with significant effect on the recipient include "feature importance" explanations. They correspond to linking client's characteristics to the recommendation, which is what feature importance techniques do. In this paper, we question the usefulness of these explanations required by law, by studying the effects of feature importance explanations on users' appropriate reliance and trust in the recommendation. In each of the studies presented below, we used SHAP [30] a post-hoc, agnostic, and widespread interpretability method as a basis to produce different explanation interfaces that vary in representation format and interactivity.

4 STUDY 1: QUALITATIVE UNDERSTANDING OF THE NEED FOR EXPLANATIONS FROM REGULATORS' AND END-USERS' PERSPECTIVES

To answer our RQ1 and RQ2, we interviewed domain experts and lay users to better understand regulatory expectations with regard to explanations.

4.1 Method

4.1.1 Prototypical Graphical Explanations. Initially, we designed an explanation interface inspired from the graphical Shapley explanations presented in [30]. However, we tried to simplify the visual elements to make them readable by non-professional users. We simplified the graph into a table, because some research on explainability showed that tables were the most interpretable representation medium for non-professional users [18]. We also added clear column titles and textual descriptions available on demand on the "input features" of the explanation, *i.e.* the client's characteristics used. We showed to participants in Study 1 a prototypical "graphical" summary of the importance of each variable on the risk of the proposal, as shown in Figure 2A. However, the arrows for each input were shaped a little differently and there was no risk scale under the different insurance plans. We improved the explanation representation based on the feedback from expert and lay participants in this study.

4.1.2 Participants and procedure. We conducted interviews with 11 participants: 6 consumer protection experts[3] and 5 end-users. The consumer protection experts were volunteers from the consumer protection section of the French regulator of banking and insurance services with whom we collaborated during this study. We refer to them below with the term "regulator". All participants had strong experience in auditing insurance providers (from 3 to more than 10 years). Their expertise and role is to verify that insurers respect "the rules intended to ensure the protection of the customers" as well as the "adequacy of the means and procedures which they implement for this purpose" and to promote fair commercial practices among industrial professionals[4]. Half of them had some experience in reviewing robo-advisors.

The novice users were volunteer doctoral students recruited through the network of the university with which the authors are affiliated. All participants received a consent form informing them of the study objectives and identified risks. All participants were volunteers, not compensated, recruited through an email describing the objective and duration of the experiment. An ethics committee was not required for this study.

Each participant took part in an individual session that lasted between 45 minutes and 1h30. Each session was divided into three

[3]Four of them were different from the 4 persons we interviewed to design the Robex algorithm.
[4]https://acpr.banque-france.fr/en/customer-protection/professionals/customer-protection-principles

parts: a semi-structured interview, a task-oriented think aloud portion and a post-study questionnaire. One researcher was present during all interviews and took detailed notes of the participants' answers and think-aloud statements. The first part of the session consisted of a semi-structured interview to explore the needs of life-insurance clients for explanations of recommendations. Structured questions varied slightly if participants were regulators or novice end-users. Regulators were asked about the role of explanations in enabling users to make an informed decision and the type of explanations, what they thought of the explanations currently offered by robo-advisors, and how to address people without financial knowledge. Novice users were asked about their financial investment recommendations, if they had any, and about what explanations they would like to receive about the recommended financial product. During the second part of the study, participants were asked to use Robex. Participants were observed by the researcher and asked to think aloud throughout their interaction with the system. Finally, participants were asked about their overall impression of the system.

4.1.3 Text analysis. We conducted an inductive [13] content analysis of the detailed notes taken by one author during the interviews with regulators and end-users. One author identified concepts and themes about the characteristics of the explanations that emerged from reading the interview notes. First, the author observed that participants talked mainly about either the explanation implementation or the explanation's purpose (notably with discussion around risk). On this basis, different themes for either explanations' format/content or explanations' purpose could be derived that encompass most of the concepts mentioned by participants. The translation from French to English was done after the final categorization.

4.2 Results

We grouped the main identified themes of the explanation requirements according to their connection to the format or content of the explanation. Through the regulator's view, we were able to gather domain perspectives that end users alone would not necessarily have provided, such as understanding the interests of different stakeholders and potential misalignment, where the vulnerability of certain users can be exploited, or the wide range of best practices seen for recommendations and explanations. Conversely, the end-users' perspective reminds us of what clients truly care about, regardless of existing regulations. While the main focus of the regulators was on the notion of risk, the main concern of the users was not as clear. For some, it was the performance of the proposed contract, for others the reliability of the robo-advisor, and for others still, the risk.

Understanding explanations' purposes through two perspectives. The regulators reported an increasing trend for automated online robo-advisors, and a lack of "good" automated explanations to support those tools. Current robo-advisors' explanations were seen as very "generic" and "nebulous" in general. One of the reasons is the use by many brokers of a third-party software to produce explanations and recommendations, over which they have little control. regulators also reported the difficulty for brokers to produce explanations with the increasing complexity of their tools: "There's too much complexity even for them." This highlights the relevance of the XAI domain to help solve real-world problems, even when the underlying recommendation system is AI but rule-based. The regulators insisted on the importance of explanations as a safeguard to inform customers about risk, taking as an example cases of overestimation of the risk for vulnerable people. Although we could group both regulator and end-user perspectives into common themes, some themes were discussed more by one group. For example, end-users expressed their need to be engaged—some felt either overwhelmed or bored by the topic. regulators talked about the need for complete information although end-users insisted on their need for simple, easy-to-digest information.

Placing the cursor between text and graphics. One of the themes we found was the need for schematic explanations on the one hand and the need for more human explanations that can answer a wide range of users' questions on the other. Two regulators very much appreciated our graphical, Shapley-based explanations, finding they had never seen something like that in the market and that it responded well to the need to link users' characteristics to the recommended product. However, many—regulators and end-users alike—indicated their need to be able to chat with a human counsellor despite the explanation. A regulator also imagined explanations could look more like a Frequently Asked Questions menu and a participant said "I can imagine a chatbot with someone behind it who can answer my questions." This led us to compare more "conversational" or more "graphical" explanations in the next study.

5 STUDY 2: DO GRAPHICAL OR DIALOGIC FEATURE-BASED EXPLANATIONS HELP LAY USERS MAKE BETTER DECISIONS?

In this large-scale study, we investigate the usefulness of simple feature importance explanations—that that can be required by law for recommender systems—to help lay users appropriately rely on life-insurance recommendations.

5.1 Study design

5.1.1 Explanations design. Based on the legal requirements for explanations and the analysis of regulators' and end-users' expressed needs, we derived the following specifications for our explanations.

What to explain?

Links between recommendation and user. We use "feature importance" explanations to address the relationship between the recommended product and the user's characteristics.

Important Definitions. As highlighted by end-users and regulators in Study 1, and by prior work [7], it is essential to give the minimal background knowledge necessary to understand the financial concepts used in the recommendations and explanations. We therefore presented definitions for all important financial concepts.

Descriptions of the effect for complex user input parameters. Robex used five user input parameters: "Your risk appetite", "Your level of financial knowledge", "the amount to invest proportionally to your total financial assets", "Your financial objective" and "The portion of your financial assets already invested". Out of those five parameters, we saw in Study 1 that the last three were more complex to interpret. For each of these concepts, we provided (1) the effect it

Table 1: Main themes emerging from the content analysis of regulators and end-users interviews, with corresponding lexical field and citations.

Explanation aspect	Regulator view	End-user view
Format and content		
Synthetic **vs.** exhaustive	short, simple, readable, "[Explanations] are a sort of synthesis", "clean and clear" **vs.** exhaustive, "Just putting a sentence *"considering this and that..."* is not enough", give links to more information, give enough documentation	simple, "Something that tells you "this is really the points you need to know""
Schematic	"schematic", "graphics and diagrams [for novice users]", "playful", "step-by-step"	"I want to see the scale of the risk, and where I'm placed on that scale"
Adapted vocabulary	"adapt vocabulary", "not too much text", "avoid financial jargon"	"use simplified language, not the language of a banker", "need to have more familiar language", "I'm not sure what a placement is"
Purpose		
Justify	link user characteristics and product, "justification", "real need of transparency" motivated by misalignment of interest between insurers and clients, prevent "scams", "what it is based on?"	"Why are you making this recommendation? What factors are you basing it on?", "I want an explanation only if there is a disagreement."
Warn	control, notify, warn, inform, "tendency to underestimate [the risk]", "Explanations are useful because there is a risk.", "the [human] advisor will not say everything", "robo-advisors don't have enough safeguards", "make them [the users] understand that there is a step to take, make them question "do I agree?""	"What are the risks?", "How much do I concretely risk losing on the 50,000 I put in?", "What can I expect in terms of risks and benefits?"
Engage users		"It looks boring", "I'll open them [the links] and probably not look at them."
Teach	enable users to have answers to their follow-up questions	"I don't know anything about that.", "I neither agree nor disagree because I don't really understand this financial concept", "I don't understand this field"

should have on the proposition—either lower or increase the risk the customer can take—(2) an indication of the magnitude of the user's input (*e.g.* "75% is a very big portion"). An example is shown in Figure 2.

In which format?

Graphical-static. The "graphical" explanation we had initially prototyped for Study 1 was improved based on participants' feedback.

Graphical-mutable. As some end-users in Study 1 expressed the need to change the parameters to know if they can trust the system, we implemented a version of the graphical explanation where user parameters were "mutable". This supports Miller's view that explanations should enable to "mutate" events [34].

Dialogic. Following feedback from end users and regulators on how textual explanations compare to human advisors', we also designed a "dialogic" explanation.It mimics a text message chat. This approach has been adopted in previous XAI work by [16, 17] for "conversational" explanations.

5.1.2 Experimental Conditions. Participants were divided into four groups corresponding to the following four different interfaces: no explanation (control group), graphical-static, graphical-mutable and dialogic. The same contextual information was delivered across all the different explanation conditions. Each of the four groups was then divided in two: one received a correct recommendation and the other a false recommendation. The objective was to compare the ability of users of different interfaces to detect a crude recommendation error.

The false recommendation was produced by altering the score-based algorithm so that the recommendation was either much too risky or really not risky enough. This was done by altering the initial user's risk score calculated by Robex by a roughly 50% change. The direction of the change was so that more-than average risk-takers were redirected to low-risk proposals and vice versa. For example, if a participant was recommended "Securimax" by the normal Robex algorithm, her risk-score would be increased artificially so as to output the "Flexiplus" recommendation. On the contrary, participants for whom the initial correct recommendation was 'the

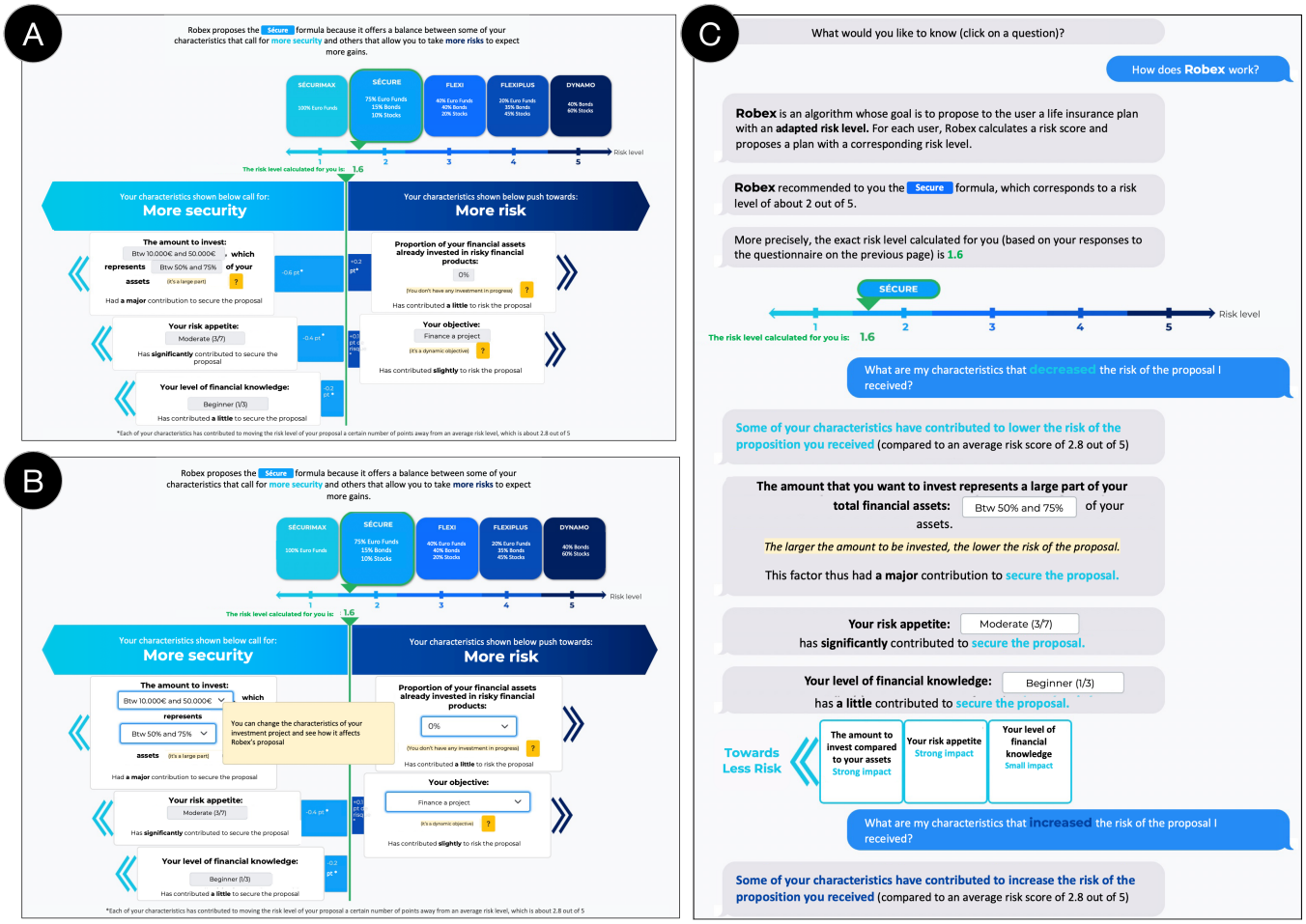


Figure 2: Explanation interfaces for each of the three conditions: A) Graphical-static: users see a graphical summary of how their characteristics impact the risk of the proposal, B) Graphical-mutable: users first see the graphical-static interface and then a pop-up message indicates they can change some of their characteristic C) Dialogic: the same information provided in the interfaces A) and B) is delivered through "sms-like" textual messages. Some graphics are added to facilitate the visualisation of the risk and of the variables decreasing and increasing the risk of the proposal. The figures are here translated to English but were shown in French to participants (cf. Figure 6 in the Appendix).

more risky 'Flexiplus" would be recommended the more conservative 'Securimax" product. For participants who initially got the "Flexi" recommendation, if their risk-score was below 12—out of a maximum score of 21—, they were redirected to "Dynamo" and for risk-scores above 12, to "Securimax".

The explanations of the false recommendation were produced in the same way as the correct recommendations, using agnostic SHAP feature importances based on the skewed Robex algorithm. As a result, the explanations for false recommendations were illogical, such as "Your risk appetite: low (1/7) contributed to increase the risk of the recommendation" *cf.* Figure 5 of the Appendix.

Participants were distributed randomly in eight different conditions as shown in Figure 3.

5.1.3 Evaluation measures. Building on prior work conducting empirical studies to evaluate XAI systems [8, 25, 29, 42], we used measures described below. Question wordings and Cronbach's alphas for grouped questionnaire items are provided in the Table 2 of the Appendix.

Reliance. Reliance was measured by asking participants if they thought the robo-advisor's recommendation was adapted to their need or not. Over-reliance occurs when the participant followed an incorrect recommendation.

Trust. Trust was measured through the five question items from the benevolence and competence aspects of McKnight's framework [32]. One item was added to measure if participants felt the need for any additional human advice.

Cognitive load. Cognitive load was measured through the mental demand and effort items of the NASA-TLX Index.

User engagement. Three user engagement question items were adapted from O'Brien *et al.*'s framework [37]. Two items were taken from the Felt Involvment (FI) category and one from the Novelty

category (NO).

Objective Understanding. Understanding of the recommendation on the one hand and understanding of the explanation on the other were measured through "test" questions. The question about the recommendation was developed by the authors relying on their knowledge of the field and discussions with experts. To measure understanding of the explanation, we used three questions to test if they understood the direction of the impact of some user inputs, as seen in prior XAI work [47].

All Cronbach's alpha's for the different sets of questions were significant, with the exception of trust for which we had to remove the question about the human advisor.

5.1.4 Procedure and participants. Figure 3 illustrates the experimental workflow used for this study. The study was approved by an academic research ethics committee. We crowdsourced participants using the platform Lucid[5]. Our goal was to target participants who might be life insurance robo-advisor users. We therefore began with a question to filter out users who were not at all interested in life-insurance. Participants were then given an overview of the study, were asked for their consent to participate in it, and went through an attention check. The two following steps in the study process replicate what we can see in existing robo-advisors: a profiling questionnaire and a following recommendation page. Participants had to go through the questionnaire, read through their user profile summary, the description of the recommendation, if applicable, an explanation of why this recommendation was made to them, and then they had to choose whether to accept or reject the proposed life-insurance plan. We also collected their qualitative feedback about explanations through a short free-text field. Finally, a two-page post-questionnaire measured their understanding, workload, trust and engagement in using Robex. The whole study lasted around 10 minutes. Participants were paid around 3€50[6] for completing the study. We randomly assigned participants to an experimental condition until we had reached a minimum of roughly 30 participants per condition. Participants who failed attention checks, took less than 5 minutes or wrote non-serious content (repeated keyboard strokes, clearly ironical or insulting content) in the free-text field were excluded. We also implemented time counters: participants could not continue to next page if a (small) minimum amount of time had not elapsed. This was to make sure that participants read through the profiling questionnaire, the recommendation and the explanation. We ended up with 32 participants in each condition.

French workers between 18 and 65 years old were recruited online through the platform Lucid. Of the study respondents that were finally included in the survey, 73% were female and 27% male—although some participants did not provide any answer to that question. 61% had an undergraduate or a graduate degree (Bachelor, Master, Doctorate and other specialized education). We cannot explain the skew towards women participants but it is possible that more male participants did not want to answer this demographic question or that our filters about the interest in life-insurance or seriousness of the responses excluded more male participants. Participants had an average financial knowledge score of 1.3 out of 5, and were therefore for the most part representative of non-expert users. Financial knowledge was measured in the pre-questionnaire through specific questions written with the help of four regulators from the French Regulation Authority of financial services (cf. Table 3 of the Appendix for the detail of the questions).

At the end of the survey, participants in the deceptive condition were informed that they had received a wrong recommendation. All participants were reminded that the financial advice presented was fictitious and non-relevant for their personal needs.

5.2 Results

For all evaluation measures, we ran a two-way ANOVA analysis with the explanation conditions and the recommendation conditions (correct or false) as the independent variables. When significant, we conducted post-hoc Tukey's HSD test for pairwise comparisons. For all measures, the assumptions for ANOVA were met: we used the Shapiro-Wilk test to check that the residuals were approximately normally distributed and the Bartlett test to verify the homogeneity of variances.

5.2.1 The no-explanation control group was more or equally likely to distinguish between good and bad advice than the explanation groups. We found a statistically significant difference in trust (p=0.001) and reliance (p=0.01) between the group that received a correct proposal and the group that received an incorrect advice for the control condition (participants who didn't receive any explanation). Yet, we sometimes didn't find such a significant statistical difference for the groups in the explanation condition. For the dialogic explanation condition, there was no statistical difference between the groups receiving a correct and an incorrect recommendation regarding trust and reliance on the advice. For the graph-mutable explanation condition, we found participants were able to differentiate their reliance on the advice between the incorrect and correct proposal (p=0.03), but not their trust. In the graphic-static explanation condition, people trusted a correct proposition significantly more than an incorrect one (p-value=0.05) and relied on the correct proposition almost significantly more (p=0.064) than on the incorrect one.

5.2.2 Dialogic explanations increase subjective trust. We found that users who were shown an incorrect recommendation and a dialogic explanations trusted significantly more the robo-advice compared to the no-explanation group (p=0.001). Further, we found that participants in the incorrect recommendation and dialogic explanation condition were almost significantly (p=0.068) more likely to rely on the incorrect robo-advice than participants in the incorrect/control condition.

5.2.3 Dialogic or graphical explanations do not improve user understanding. The different explanation formats did not improve users' understanding of the recommendation and more specifically its risk —question one out of three on the recommendation understanding (cf. Table 2 in the Appendix). Based on the graphs in Figure 4, there appears to be a tendency for graphical-mutable explanations to lead to better understanding of the recommendation than other conditions, but the effect was not significant (p=0.1). Further, the level of understanding of the explanations was comparable across the different explanation conditions. However, people in the deceptive

[5]https://lucid.co/

[6]Lucid goes through several suppliers to gather participants. Each supplier receives 3.50€ for each study completed, takes a commission and pays the rest to the participant.

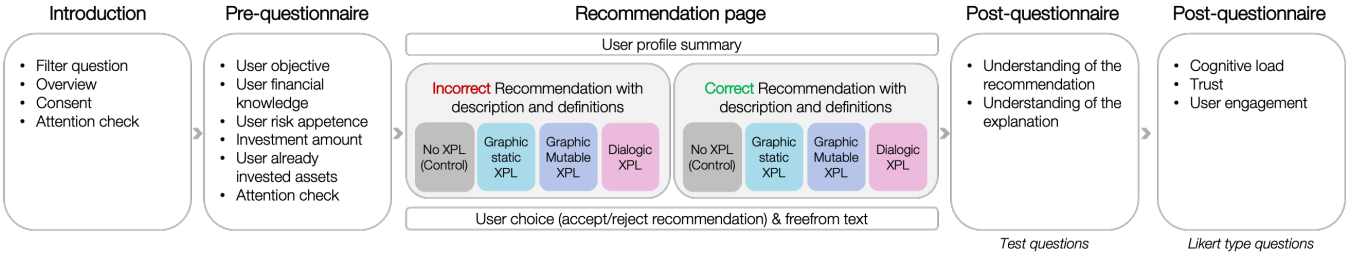
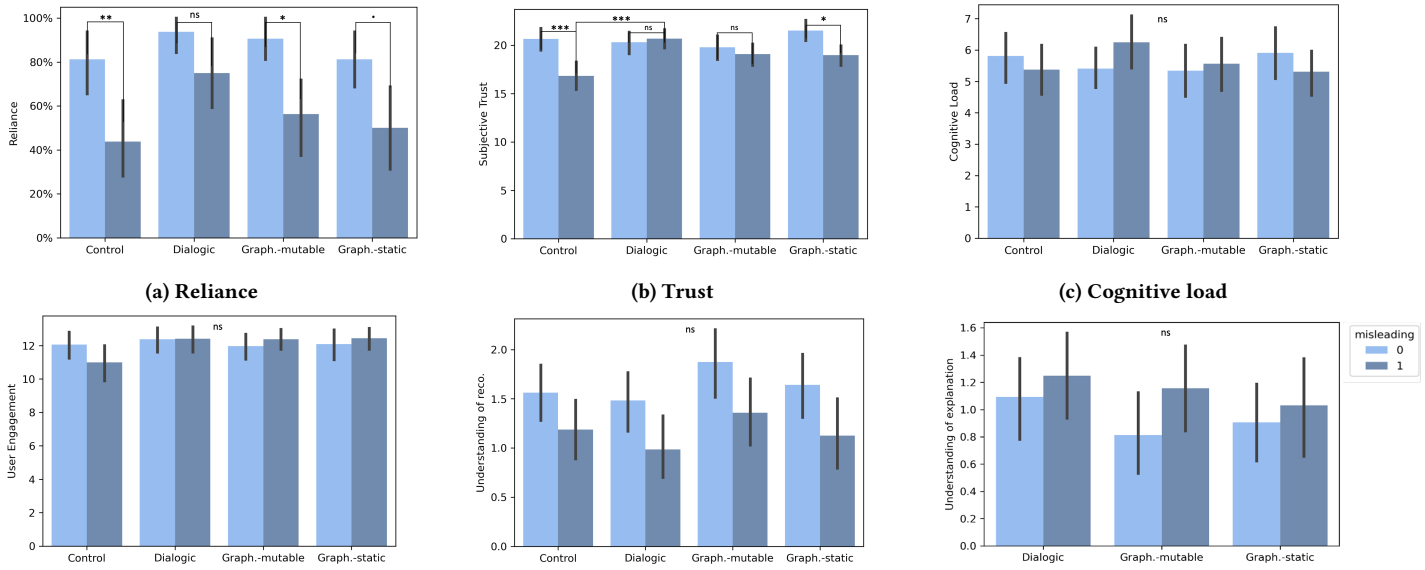


Figure 3: The workflow of our quantitative experiments. The profiling questionnaire is used to produce a personalized recommendation of a life-insurance contract. Clients can review the recommendation, the explanation and then decide to follow the recommendation or not. This decision is used to measure users' "reliance" on the explainable robo-advisor.



(a) Reliance **(b) Trust** **(c) Cognitive load**

(d) User engagement **(e) Understanding of the recommendation** **(f) Understanding of the explanation**

Figure 4: Results for Study 2. Vertical lines represent the 95% confidence interval. Asterisks and dots indicate the statistical significance of the results: *** p-value≤0.001, ** p-value≤0.01, * p-value≤0.05, • p-value≤0.07, "ns" non significant.

conditions were significantly less likely to understand the characteristics of the recommendation and the explanations (p=0.001)—we performed a one-way ANOVA with just the recommendation condition (correct or false) as the independent variable. This evidences that people are less likely to understand a recommendation that is not suited to their needs, or that they did not expect.

5.2.4 No effect of explanations on cognitive load and user engagement. We do not find any statistically significant effect for the different explanation conditions on users' subjective cognitive load and user engagement. This finding contradicts other work on the cognitive cost of explanation [49]. Perhaps this is the case here because understanding financial recommendations is already cognitively demanding enough due to the complexity of the field, and the cost of adding explanations is negligible in comparison—average perceived cognitive workload for using the robo-advisor was 5.6 out of 10.

6 LIMITATIONS

This work has some limitations. First, the content analysis in Study 1 was performed based on the detailed notes that one author took during the interviews, which may have limited the amount and breadth of captured input from participants. In addition, the non-expert participants from the qualitative study were graduate students, who represent a very specific sample of non-expert users. One of the limitations in our implementation of ecological interface design is that we used a simplified and fictional life-insurance robo-advisor. Some factors such as time horizon, detailed descriptions of the funds, of their historical performances and the costs of each contract were not taken into account. We did this to simplify the building of the tool, and also because we felt adding costs and performances might have diverted participants' focus from the risk of the proposals, which is the most critical information for users to understand according to regulators and the spirit of the legislation. Future work

could explore similar research questions with a real robo-advisor. Additionally, one of the main limitations of crowd-sourcing participants in Study 2 is that they might lack the mental engagement or involvement with the subject. To increase participant engagement, we let them answer the survey with their own profile, instead of presenting a predefined profile for all participants. We verified that the type of recommendation did not have a significant impact on our measures. Additionally, we implemented a question to filter out users completely uninterested in life-insurance, attention checks, text fields and time counters to filter out non-serious participants. Nevertheless, it is possible that the participants in our study were not representative of a real user of a real life-insurance robo-advisor. Also, the participants in our study were also mainly women (73%).

7 DISCUSSION AND FUTURE WORK

Dialogic vs. Graphical explanations. According to Miller [34], explanations are best provided through a social process, *i.e.* a conversation, because it matches the way humans explain things. In fact, "dialogic" explanations have been favorably presented in the XAI literature, with [17] presenting how dialogic management systems can respond to users' questions about a hotel recommender system, or [16] showing how conversational explanations can be useful for criminal investigators. While the benefits of dialogic explanations might be real regarding user satisfaction and explanation usefulness in some contexts [16, 17], our results, in turn, shed light on a downside of "dialogic" explanations for impactful AI-based decisions: over-reliance. It is possible that either the "humanness" of the dialogic explanation we presented, or the familiarity of users with chats, made them more inclined to accept robo-advice. In fact, some people might see the anthropomorphisation of systems as suspicious. One of our end-user participants in the pilot Study said that "It's quite a lot of anthropomorphization". This is consistent with the study by Hepenstal *et al.* [16] in which participants were uncomfortable with the humanness of the XAI agent and wanted to have it clear that they were not talking to a real person. Our findings also qualify Szymanski *et al.*'s results [47] according to which participants prefer graphical explanations but understand textual explanations better. The authors further advance that hybrid textual and graphical formats could improve both user satisfaction and understanding. Our study qualifies this result by showing that users made less mistakes with graphical formats which presented small amounts of text than with dialogic formats with small amounts of graphical visualizations. This contrasts with Szymanski *et al.*'s finding that text is better understood—however the textual explanations in this work were much shorter. Perhaps the brevity and the synthetic aspect of our graphic explanations compared to the dialogic explanations were instrumental in improving users' appropriate reliance.

Legal requirements for feature-based explanations. In this study, we showed how legal requirements to "motivate" investment advice based on client's features may take shape using a classical XAI method (SHAP) and various explanation representations. We further found that the legal sub-objectives of the explanation that we defined in Section 3.1.3 to help users make "fully informed" decisions were not fully achieved. Users were not better able to 1) appropriately rely on the recommendation, 2) understand the recommendation or 3) appropriately calibrate their trust in the robo-advisor compared to the control condition. As noted in Section 3.1.3, the objective of the law requiring insurance intermediaries to specify in writing "the reasons for the appropriateness of the proposed contract" is also to discipline brokers by making non-objective, self-interested, recommendations more visible and punishable. Feature-based explanations are therefore not useless, because they at least serve the purpose of disciplining insurance intermediaries by forcing them to show how the proposed product corresponds to the customer's risk profile. However, our work changes the perspective on the benefit of explanations for customers' understanding and reliance. Explanations are not always "all good", they must be designed so that over-reliance effects are mitigated. If the explanation formats we presented could not meet the legal objectives we highlighted, future work could address how to design explanations that are cognitively engaging for lay-users. Buçinca *et al.* designed cognitive forcing functions, but these were perceived as friction by the users. Melsion *et al.* [33] designed "quiz" explanations by asking users—in this case children—what they thought were the most important characteristics for an AI to predict gender. The use of such gamified explanations could improve learning in a specific domain without sacrificing user satisfaction.

8 CONCLUSION

In this paper, we carried out a qualitative study to understand what end-users and consumer protection experts—regulators—say about feature-based explanation requirements. We then presented the results of a large-scale study to investigate if different formats of feature-based explanations help novice users appropriately rely on, trust and understand recommendations of life-insurance plans. We found that providing feature-based explanations did not significantly improve users' understanding of the recommendation, or lead to more accurate reliance on the tool's recommendation compared to having no explanation at all. We also found that explanations provided in a dialogic format, where users can choose a question and get chatbot-like text answers, increased users' trust in the robo-advisor and did not significantly improve user understanding. This led us to conclude that graphical formats could be better suited to inform clients. This leaves us in a quite unsatisfactory state of affairs where the obligation to inform clients does not fulfill its promises to empower users in making better decisions. We highlighted promising future leads to address this challenge. Finally, we hope our work may encourage researchers to investigate how legal explainability requirements may take shape, and how to address the problem of informing non experts in complex domains.

ACKNOWLEDGMENTS

This research is sponsored by the Agence Nationale de la Recherche (ANR) through the grant ANR-20-CHIA-0023-01 and by the Af2i (Association française des investisseurs institutionels) through the Young Researcher Award attributed to Astrid Bertrand. We thank Olivier Fliche, Christine Saidani, Laurent Dupont, and all the participants from the ACPR and Télécom Paris for their helpful guidance, comments and for making this project possible.

REFERENCES

[1] Ramnath Balasubramanian, Ari Chester, and Nick Milinkovich. 2020. *Rewriting the rules: Digital and AI-powered underwriting in life insurance.* Consultancy Report. McKinsey & Company. https://www.mckinsey.com/industries/financial-services/our-insights/rewriting-the-rules-digital-and-ai-powered-underwriting-in-life-insurance

[2] Ramnath Balasubramanian, Ari Libarikian, and Doug McElhaney. 2021. *Insurance 2030—The impact of AI on the future of insurance.* Technical Report. McKinsey & Company. https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance

[3] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2021. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 0, 0 (2021), 1–29. https://doi.org/10.1080/12460125.2021.1958505 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/12460125.2021.1958505.

[4] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22).* Association for Computing Machinery, New York, NY, USA, 78–91. https://doi.org/10.1145/3514094.3534164

[5] Astrid Bertrand, Winston Maxwell, and Xavier Vamparys. 2021. Do AI-based anti-money laundering (AML) systems violate European fundamental rights? *International Data Privacy Law* (April 2021). https://doi.org/10.1093/idpl/ipab010

[6] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2021. Legal Requirements on Explainability in Machine Learning. *Artificial Intelligence and Law* 29, 2 (2021), 149–169. https://doi.org/10.1007/s10506-020-09270-4 Publisher: Springer Verlag.

[7] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces.* ACM, Helsinki Finland, 807–819. https://doi.org/10.1145/3490099.3511139

[8] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 188:1–188:21. https://doi.org/10.1145/3449287

[9] Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zytek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni. 2022. VBridge: Connecting the Dots Between Features and Data to Explain Healthcare Models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 378–388. https://doi.org/10.1109/TVCG.2021.3114836 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[10] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[11] Jonathan Dodge, Andrew A. Anderson, Matthew Olson, Rupika Dikkala, and Margaret Burnett. 2022. How Do People Rank Multiple Mutant Agents?. In *27th International Conference on Intelligent User Interfaces (IUI '22).* Association for Computing Machinery, New York, NY, USA, 191–211. https://doi.org/10.1145/3490099.3511115

[12] Malin Eiband, Daniel Buschek, and Heinrich Hussmann. 2021. How to Support Users in Understanding Intelligent Systems? Structuring the Discussion. *arXiv:2001.08301 [cs]* (Feb. 2021). http://arxiv.org/abs/2001.08301 arXiv: 2001.08301.

[13] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of Advanced Nursing* 62, 1 (2008), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2648.2007.04569.x.

[14] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19).* Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[15] Juliana J. Ferreira and Mateus S. Monteiro. 2020. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments (Lecture Notes in Computer Science),* Aaron Marcus and Elizabeth Rosenzweig (Eds.). Springer International Publishing, Cham, 56–73. https://doi.org/10.1007/978-3-030-49760-6_4

[16] Sam Hepenstal, Leishi Zhang, Neesha Kodagoda, and B. l. william Wong. 2021. Developing Conversational Agents for Use in Criminal Investigations. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Dec. 2021), 1–35. https://doi.org/10.1145/3444369

[17] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *CUI 2021 - 3rd Conference on Conversational User Interfaces (CUI '21).* Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3469595.3469596

[18] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (April 2011), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

[19] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445385

[20] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Transactions on Computing for Healthcare* 1, 1 (March 2020), 6:1–6:20. https://doi.org/10.1145/3344258

[21] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not Enough, Learn to Criticize! Criticism for Interpretability. (2016), 11.

[22] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). https://doi.org/10.48550/arXiv.1711.11279 arXiv:1711.11279 [stat].

[23] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. *arXiv:2002.11097 [cs, stat]* (June 2020). http://arxiv.org/abs/2002.11097 arXiv: 2002.11097.

[24] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2019. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 299–309. https://doi.org/10.1109/TVCG.2018.2865027 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[25] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. https://doi.org/10.48550/arXiv.2112.11471 arXiv:2112.11471 [cs].

[26] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021). https://doi.org/10.1016/j.artint.2021.103473

[27] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[28] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing (UbiComp '09).* Association for Computing Machinery, New York, NY, USA, 195–204. https://doi.org/10.1145/1620545.1620576

[29] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 408:1–408:45. https://doi.org/10.1145/3479552

[30] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[31] YOKOI-ARAI Mamiko. 2020. *The Impact of Big Data and Artificial Intelligence (AI) in the Insurance Sector.* Technical Report. OECD. http://www.oecd.org/finance/Impact-Big-Data-AI-in-the-Insurance-Sector.htm

[32] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (Sept. 2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81 Publisher: INFORMS.

[33] Gaspar Isaac Melsión, Ilaria Torre, Eva Vidal, and Iolanda Leite. 2021. Using Explainability to Help Children UnderstandGender Bias in AI. In *Interaction Design and Children.* ACM, Athens Greece, 87–99. https://doi.org/10.1145/3459990.3460719

[34] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[35] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv:1811.11839 [cs]* (Aug. 2020). http://arxiv.org/abs/1811.11839 arXiv: 1811.11839.

[36] Mohammad Naiseh, Reem S. Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through Friction: An Approach for Calibrating Trust in Explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*. 1–5. https://doi.org/10.1109/BESC53957.2021.9635271

[37] Heather O'Brien and Paul Cairns. 2015. An empirical evaluation of the User Engagement Scale (UES) in online news environments. *Information Processing & Management* 51, 4 (July 2015), 413–427. https://doi.org/10.1016/j.ipm.2015.03.003

[38] Aimee Prawitz, E. Thomas Garman, Benoit Sorhaindo, Barbara O'Neill, Jinhee Kim, and Patricia Drentea. 2006. Incharge Financial Distress/Financial Well-Being Scale: Development, Administration, and Score Interpretation. https://papers.ssrn.com/abstract=2239338

[39] Juan Rebanal, Jordan Combitsis, Yuqi Tang, and Xiang 'Anthony' Chen. 2021. XAlgo: a Design Probe of Explaining Algorithms' Internal States via Question-Answering. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 329–339. https://doi.org/10.1145/3397481.3450676

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[41] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[42] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (Feb. 2021), 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

[43] Auste Simkute, Ewa Luger, Mike Evans, and Rhianne Jones. 2020. Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference (DIS' 20 Companion)*. Association for Computing Machinery, New York, NY, USA, 263–268. https://doi.org/10.1145/3393914.3395862

[44] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 56–67. https://doi.org/10.1145/3351095.3372870

[45] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 212–228. https://doi.org/10.1145/3490099.3511119

[46] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 74. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445088

[47] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. https://doi.org/10.1145/3397481.3450662

[48] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *arXiv:1806.07552 [cs]* (June 2018). http://arxiv.org/abs/1806.07552 arXiv: 1806.07552.

[49] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. http://arxiv.org/abs/2212.06823 arXiv:2212.06823 [cs].

[50] Kim J. Vicente. 2002. Ecological Interface Design: Progress and Challenges. *Human Factors* 44, 1 (March 2002), 62–78. https://doi.org/10.1518/0018720024494829 Publisher: SAGE Publications Inc.

[51] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

APPENDIX

Table 2: Question used for measuring different metrics with Cronbach alphas (translated from French to English).

Measure	Questions with [possible responses]	Cronbach's alpha
Understanding of recommendation	What is your estimate of the euro fund percentage in the proposal that was made to you? [Several proposals]	NA
	On a scale of 1 to 5 (5 being the most risky), how risky do you think the Robex proposal is?	
	What is special about a euro fund? [it offers a high expectation of gains for a high risk of loss, it is mostly composed of actions, it is guaranteed by the insurer, I do not know]	
Understanding of explanation	Of your characteristics and goals, which factor weighed the most in the proposal the algorithm offered you? [Several proposals]	NA
	How did the proportion of your financial assets already invested in risky financial products, which is for you ... , impacted the risk of proposal made by Robex? [Increase / decrease / neutral]	
	How did your investment objective, which is ... impacted the risk of the proposal made by Robex?	
Trust-Benevolence	I think Robex is acting in my best interest	0.854
	Robex wants to understand my needs and preferences	
Trust-Competence	Robex is skilled and effective in providing life insurance recommendations	0.878
	Robex has the expertise to understand my needs and preferences	
	Robex is fulfilling its role as a life insurance advisor very well	
Trust-Other (not used)	I would need a human advisor to help me choose a life insurance plan	Not used
User engagement	I felt involved in my task of choosing a life insurance plan	0.818
	The content of the life insurance recommendation site has attracted my curiosity	
	I was interested in the experience	
Cognitive load	I found it mentally demanding to read and understand the proposed life insurance formula	0.829
	I had to make an effort to read and understand the proposed life insurance formula	

Table 3: Question used in the pre-questionnaire for measuring users' personal characteristics (translated from French to English).

Measure	Questions with [possible answers]
Objective	What would be the main objective of your investment? [Make my savings grow, Finance a project, Finance my retirement, Pass on my assets, Protect my savings]
Amount to be invested	How much would you like to invest? [Less than 5000€, Between 5000€ and 10 000€, Between 10000€ and 50000€, More than 50000€]
	This amount represents what percentage of your total financial assets (excluding your home)? [Less than 5%, Between 5% and 25%, Between 25% and 50%, Between 50% and 75%, More than 75%]
Percentage of assets already invested	Have you already invested in a financial product with a risk of capital loss? If so, how much of your total financial assets do these financial products represent? [Less than 5%, Between 5% and 25%, Between 25% and 50%, Between 50% and 75%, More than 75%]
Risk appetite	Which of the following statements is closest to the level of financial risk you are willing to take when saving or investing? [Take significant financial risk hoping for significant returns, Take above average financial risk hoping for above average returns, Take average financial risk hoping for average returns, I do not wish to take any financial risk]
	For the next three sentences, please indicate the likelihood that you would engage in the specified behavior if you were in the situation described "Investing 10% of your annual income in an investment consisting of securities issued by the European Union" [Very unlikely, Somewhat unlikely, Uncertain, Somewhat likely, Very likely]
	"Investing 5% of your annual income in highly speculative securities" [Very unlikely, Somewhat unlikely, Uncertain, Somewhat likely, Very likely]
	"Investing 10% of your annual income in a new business" [Very unlikely, Somewhat unlikely, Uncertain, Somewhat likely, Very likely]
Financial knowledge and experience	Have you ever subscribed to a life insurance contract? [Yes, No]
	Have you ever invested in a financial product with a risk of capital loss (e.g. PEA (Plan d'Epargne en Actions), multi-support life insurance contract, securities account, crypto assets, investment funds...)? [Yes, No]
	A high expectation of gains implies a high risk of capital loss. [True, False]
	A real estate fund (SCPI or OPCI) is a fund with guaranteed capital. [True, False]
	The capital invested in a life insurance plan is blocked for 8 years. [True, False] The capital invested in life insurance units of account is subject to a risk of capital loss. [True, False]

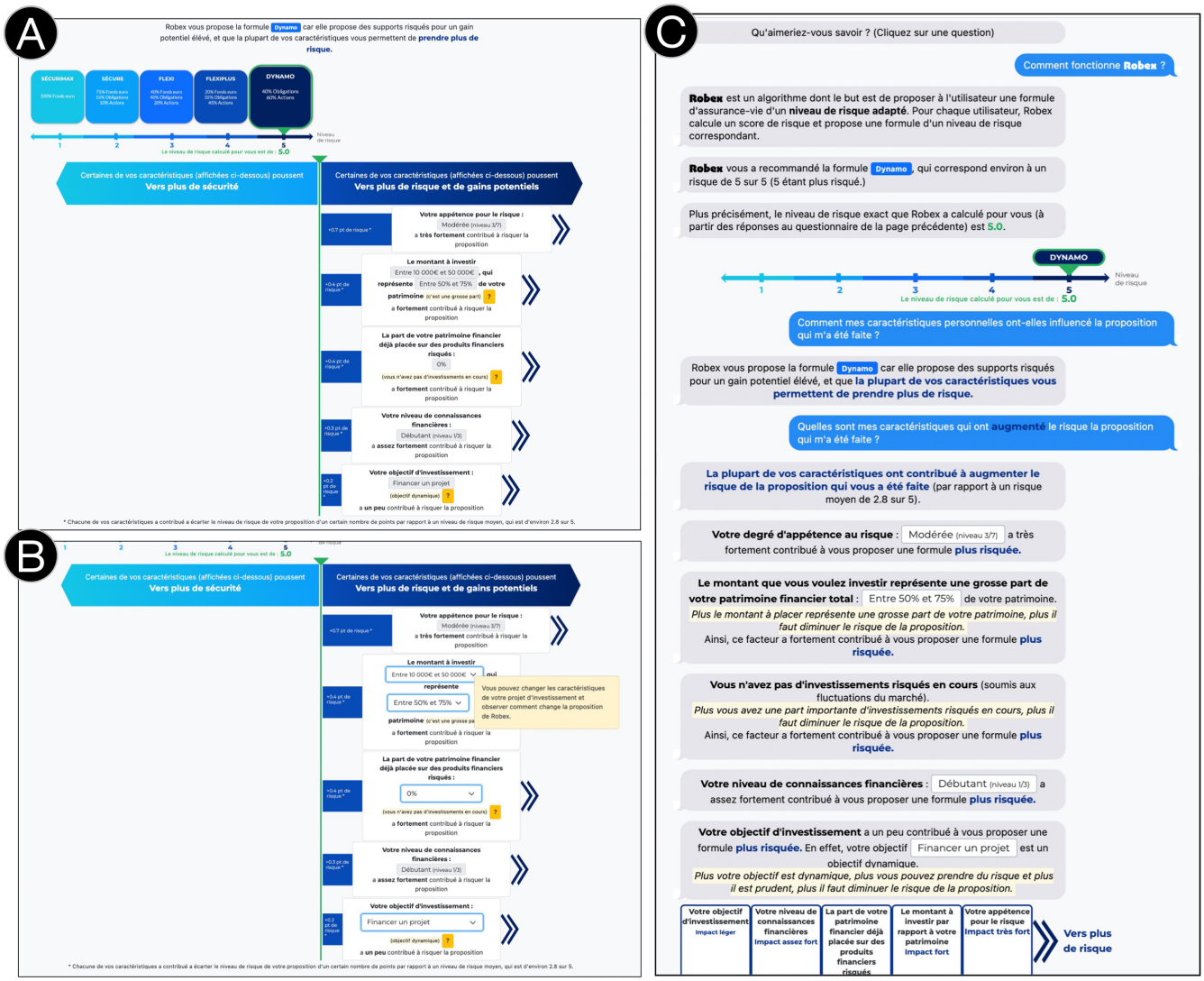


Figure 5: Explanation interfaces examples for an incorrect recommendation for each of the three conditions: A) Graphical-static B) Graphical-mutable C) Dialogic. The correct user profile in this case would have been "Secure", but the skewed Robex algorithm outputs "Dynamo". Explanations are in French, as shown to participants.

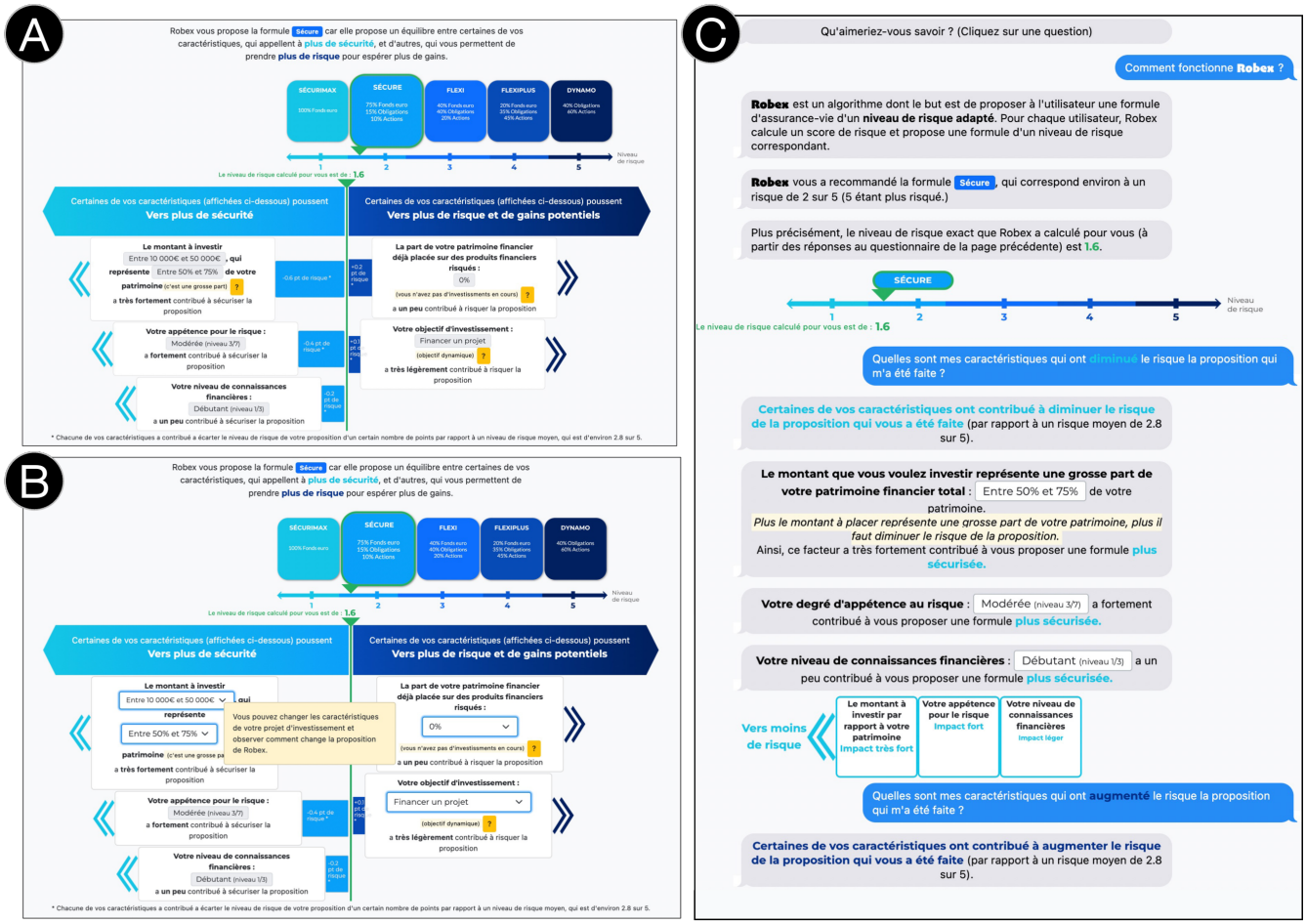


Figure 6: The original, French version of Figure 2 that shows the three explanation conditions for participants who received a correct recommendation.